The selection of optimal discriminant procedures for discrete data.

LACK, Hans Nicholas.

The Selection of Optimal Discriminant Procedures for Discrete Data

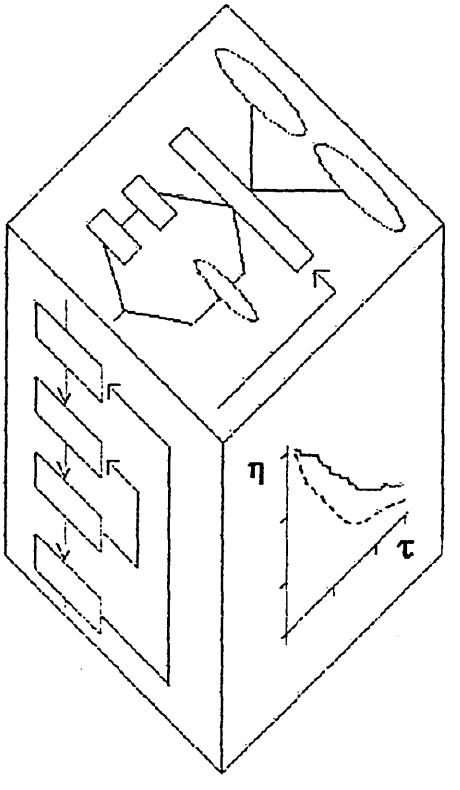
Hans-Nicholas Lack

A thesis submitted in partial fulfilment of the requirements of
Sheffield Hallam University
for the degree of Doctor of Philosophy

April 1996

Collaborating Organisation: Kassenärztliche Vereinigung Bayerns

The Selection of Optimal Discriminant Procedures for Discrete Data



Hans-Nicholas Lack

Munich, 17-04-1996

A B S T R A C T

The statistical literature contains a wide variety of reports on procedures for discriminant analysis. They range from the classical linear discriminant and logistic model at one end of the spectrum through to the recent recursive partitioning and neural network based procedures in the field of pattern recognition at the other extreme. By contrast the literature offers little advice concerning choice of procedure especially for discrete data. This thesis therefore addresses the problem of selection of optimal discriminant procedures for discrete data.

The problem is approached by identifying key determinants of procedure choice such as prior information about the data, sample size and the performance expected of the procedure. Two new ways of assessing performance of a discriminant are suggested for the discrete data situation. The first of these, the η-criterion, is a weighted sum of posteriors for correctly allocated and misallocated objects. The second consists of analysing performance in relation to the distribution of relative differences between the two largest posteriors. A selection tree is constructed on the basis of these two approaches.

The results indicate that the η-criterion exhibits low variance as well as low bias but also has the ability to differentiate better than the customary error rate. The use of classification thresholds proves particularly useful in the detection of optimal procedures for discrete data. A structured approach using a selection tree is demonstrated and evaluated.

Integration of the developed techniques into statistical software packages is recommended.

Declaration

I, Hans Nicholas Lack, hereby declare that I have completed the following thesis entirely on my own.

List of publications

The following lists publications within the period of the research for MPhil/PhD. Other related publications are listed together with other references in chapter 19.

Lack, H.N. (1992). "Qualitätsindikatoren - Entwicklung einer geeigneten Darstellung", <Indicators of quality - development of a suitable form of presentation>, in: *BPE Jahresbericht 1991*, Kommission für Perinatologie und Neonatologie der Bayerischen Landesärztekammer und der Kassenärztlichen Vereinigung Bayerns <annual report of Bavarian Perinatal Survey Steering Committee>

Lack, H.N. (1993). "Schwangerenvorsorge", <Antenatal care>, In *BPE Jahresbericht 1993*, Kommission für Perinatologie und Neonatologie der Bayerischen Landesärztekammer und der Kassenärztlichen Vereinigung Bayerns <annual report of Bavarian Perinatal Survey Steering Committee>

Lack, H.N. (1994). "Läβt sich eine drohende Frühgeburt vorhersagen ?", <Are imminent premature deliveries predictable ?>, In *BPE Jahresbericht 1993*, Kommission für Perinatologie und Neonatologie der Bayerischen Landesärztekammer und der Kassenärztlichen Vereinigung Bayerns <annual report of Bavarian Perinatal Survey Steering Committee>

Acknowledgements

I would like to express my sincerest thanks to Mike Grimsley, Director of Studies, for his continued support especially in times when progress seemed particularly difficult but also for repeated reading of the two final draft versions of the text which lead to numerous suggestions on presentation and continuity of style. Next I would like to extend my thanks towards my supervisors, Professor W. G. Gilchrist and Dr. R.J. Gadsden for their useful comments on the η-criterion and on classification thresholds as well as suggestions for some of the graphical illustrations.

I would like to express my thanks to my ever ready father, Mr. Ernest N. H. Lack who taught me that properly done jobs are worthwhile endeavours and who helped me especially in the final hectic phases of draft preparation with typing and proof reading in several locations. As such this also turned out to be a continued source of encouragement without which everything would have been even more of a challenge.

Next I would like to thank my employers in their capacity of collaborating institutions over the past seven years: the pharmaceutical company, Dr. Karl Thomae, Biberach, Germany and especially the union of doctors within the mandatory health insurance scheme in Bavaria, the Kassenärztliche Vereinigung Bayerns. They helped me considerably throughout by being understanding with respect to regular but sudden "holiday" requests and therefore giving continued support to the research.

Last, but certainly not least, I would like to thank my dear wife, Judy, for bearing with me over several strenuous years in which we shared only little time together. It was a time, however, that now makes us both look forward all the more to proper holidays again together with our sons John-Certus and John-Christopher.

List of contents

Part III - Method

Part IV - Results

Part V - Discussion

Appendices

List of figures

List of tables

Chapter 1 - Summary

"The problem of discriminant analysis may be subdivided into three categories depending upon the degree of information available about the class specific distributions F_i. These are: (1) F_i completely known which is rarely the case, (2) the usual case in which the F_i are known to belong to some parametric class of distributions so that $F_i(x) = F_i(x,\theta_i)$ where the distributions F_i are now known but θ_i are unknown and require estimation and, finally, (3) a similarly not unrealistic situation in which the F_i are completely unknown. The basic problem in discriminant analysis may then be described as finding suitable means for estimating the class specific densities, $\hat{f}_i(x|\theta_i)$, and then using these to derive allocation rules".

1.1 Introduction

It is no coincidence that the above quote dates back over *40* years which may surprise in the context of the present thesis. However it has been selected because few authors have since been able to state the discriminant problem as succinctly as Fix and Hodges did in 1951[1]. These authors addressed the problem of estimating unknown probability distributions which laid the foundations for a technique later to become known as *nearest neighbour* procedure. They also used these density estimates to obtain allocation rules for discriminating among different populations.

Unlike many other areas in statistics the subject of discriminant analysis - not necessarily restricted to discrete data - draws simultaneously on several basic statistical techniques. These include estimation theory, distributional theory, sampling theory and decision theory. Thus any more than cursory treatment of discriminant

---

[1] Fix, E. and Hodges, J.L. (1951). "*Non-parametric discriminant analysis*", United States Air Force School of Aviation Medicine, Project Number 21-49-004, Reports. 4 & 11

analysis claiming to be exhaustive will tend to be expansive. The recent book by McLachlan (1992) - frequently referred to as *the standard text on discriminant analysis* - is a good case in point. Yet even his comprehensive treatment including a bibliography of more than *1200* references only partially deals with the recent recursive partitioning procedures and gives little structural help in selecting optimal procedures. The present thesis on discriminant analysis is, therefore, restricted to discriminant analysis specifically applied to discrete datasets. Emphasis has been placed on the mechanics of deriving a structured approach to selection of optimal procedures rather than on the issues of missing data estimation and selection or transformation of variables.

The work presented in the following deals with the situation where the data are exclusively of a *discrete* nature, where prior information about the data distributions is only partially given, scant or not available at all and where a choice has to be made among a range of discriminant procedures not necessarily all specifically designed for discrete data.

In the published literature there are not many guides to selection of discriminant procedures. This is especially so in the field of discrete data where often no data model exists thus making density estimation a crucial step. A further difficulty arises because for datasets with few discrete multivariate states, estimates of the misallocation error will also tend to be unstable. As these depend on the posterior probabilities of population membership the problem is thus reduced to finding realistic estimates of the posteriors.

This problem is addressed and also extended to an investigation of the entire posterior distribution. The results show that extracting information from the posteriors readily lends itself to constructing useful criteria for performance assessment of a discriminant

procedure. Two new tools are introduced for this purpose. Firstly, a variable classification threshold is constructed, and in order to qualify for allocation an object's posterior probability must exceed this local threshold. Secondly, the distribution of relative differences between the first two largest posteriors is derived empirically and then used as an additional evaluator of the performance of a discriminant rule. Next the crucial determinants of procedure choice are identified and based upon these a suitable selection tree is constructed. A structured approach to procedure selection is developed.

The results from applying this approach to a range of "real data" taken from the published literature as well as results obtained from artificially generated data using Monte Carlo techniques indicate that the suggested performance criterion may suitably augment the common misallocation error. The computational method is straightforward and readily implemented on a small microcomputer using assembled *FORTRAN* code. The inputs required for the programs are the posterior probabilities of population membership which can also be supplied by most professional statistical software. Possible extensions of the suggested method to datasets with continuous or mixed data structures are also straightforward. Artificial datasets, expectation, bias and variance of performance estimators as well as estimates of the empirical distribution of the relative difference of maximal posteriors were generated using bootstrap techniques. The procedure selection approach is demonstrated using real and artificial data. The structured approach and use of the new performance criteria enables choice of reliable procedures.

1.2 Layout of thesis

The entire thesis comprises all the parts shown in fig. 1.2-1

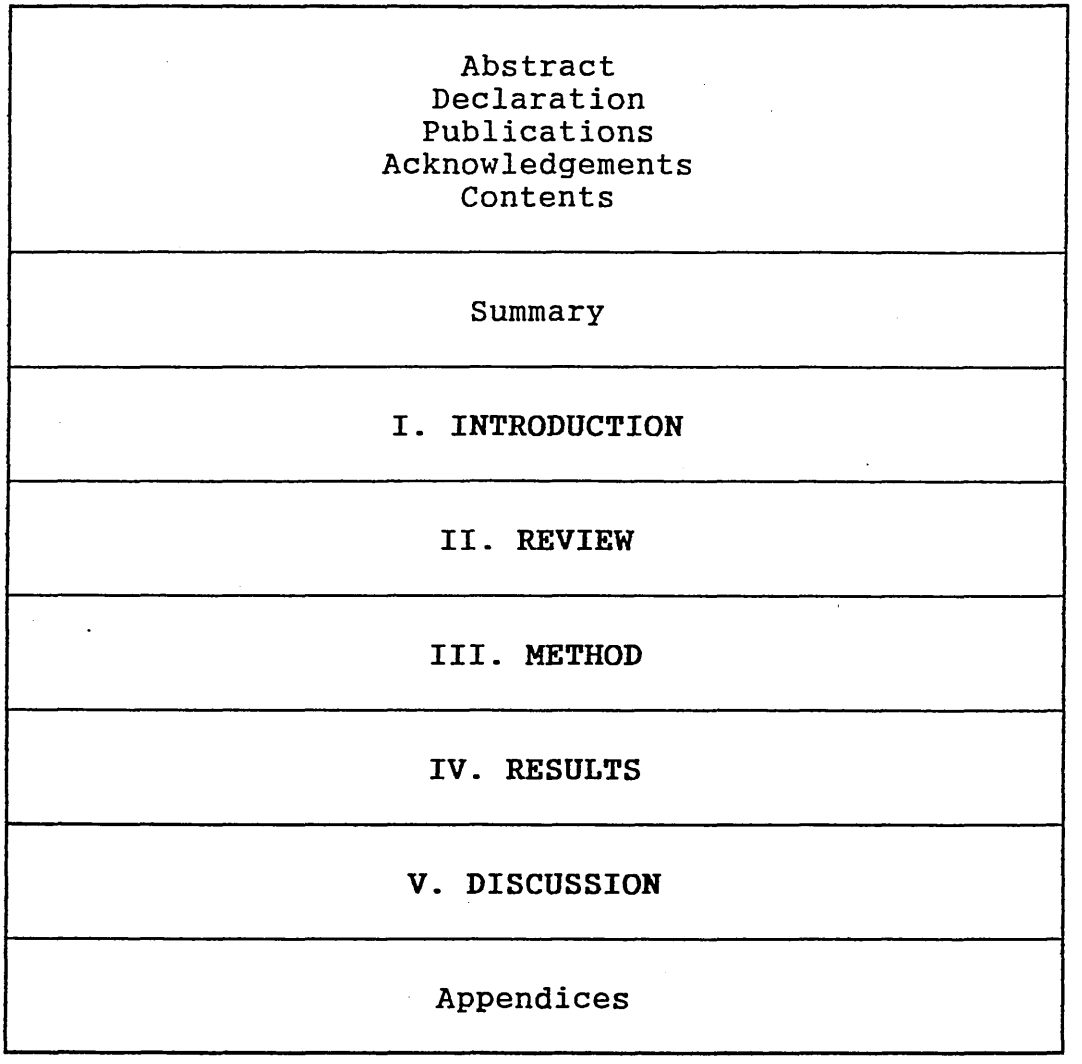


Figure 1.2-1: Layout of thesis

The main body of the thesis breaks down further into *5* major parts following the classical layout: (I) Introduction, (II) Review, (III) Method, (IV) Results and (V) Discussion. These parts are emphasised in bold type in Fig.1-1. The following briefly describes parts I to V.

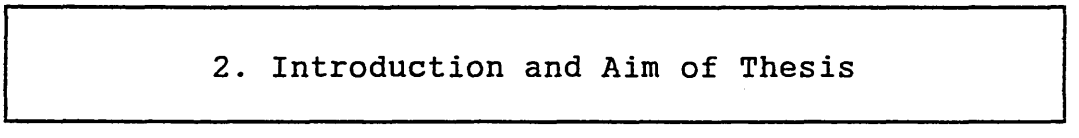


Figure 1.2-2: Part I: Introduction

A non technical introduction to the subject is given using a series of examples which illustrate a variety of data situations. The examples show all of the key issues addressed later: the data dimension ranging from continuous to discrete, linear and curvilinear separation lines, the effects of scaling of variables, overfitting and parsimonious models, sampling effects, pattern recognition, interactions between variables, partitioning techniques and selective discrimination. This part ends with a specification of the research aims.

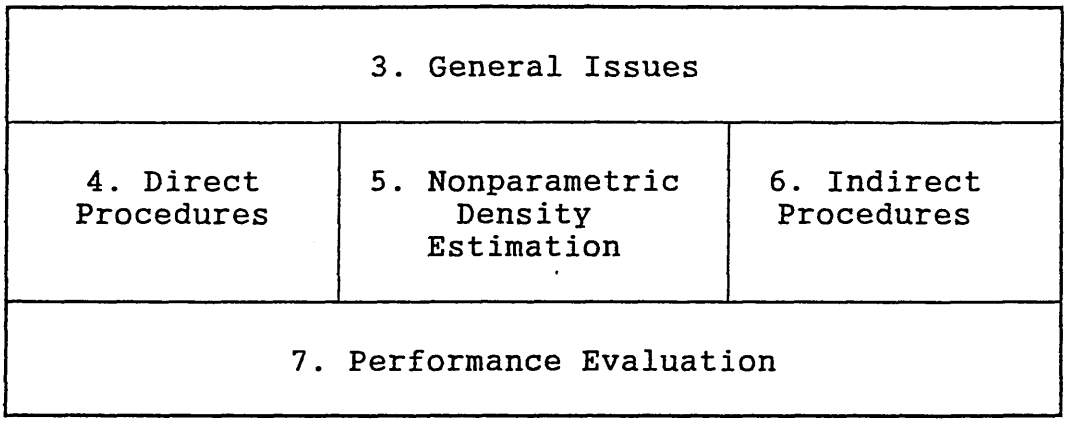


Figure 1.2-3: Part II: Review

This part begins with a coverage of general issues in discriminant analysis (chapter 3). Next direct discriminant procedures that use the posteriors explicitly in allocation are treated separately (chapter 4) from indirect procedures that use other techniques such as interpopulation distances (chapter 6). As the distribution of posterior probabilities across discrete datasets depends on the respective population specific densities their estimation is vital to the success of a discriminant procedure. In the case of discrete data a parametric model may not always exist. An entire chapter has therefore been devoted to the review of nonparametric density estimation (chapter 5). The review ends with a chapter on the evaluation of performance of a given discriminant procedure (Chapter 7).

8. Performance Criteria	9. Classification Thresholds
10. Technical Issues	
11. Data Sets	
12. Construction of Selection Rules	

Figure 1.2-4: Part III: Method

The review (part II) indicates that improvements on performance evaluation are needed for establishing a useful selection guide. Hence the first two chapters deal with construction of suitable performance criteria (chapter 8) and classification thresholds (chapter 9) respectively. The use of classification thresholds, i.e. the specification of a minimum threshold to be exceeded by the posterior probability, is new and non standard practice. For this reason it is treated in a chapter on its own. The next two chapters describe necessary technical refinements and adjustments to procedures (chapter 10) and also give an overview of the real and artificial datasets used in the examples (chapter 11). Finally the construction of selection rules is developed in chapter 12.

13. Analysis of Performance Criteria	14. Analysis of Classification Thresholds
15. Application of Selection Rules	

Figure 1.2-5: Part IV: Results

The results are threefold and are presented in separate chapters: the analysis of performance criteria (chapter 13), the analysis of classification thresholds (chapter 14) and application of the selection rules (chapter 15).

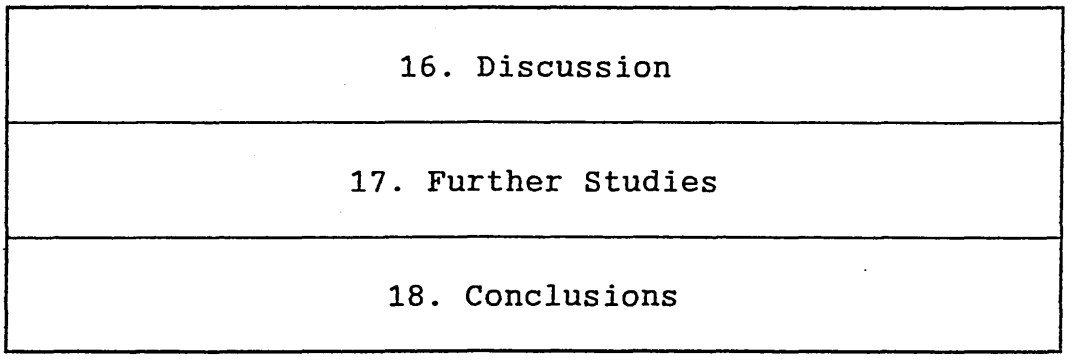


Figure 1.2-6: Part V: Discussion

The final part begins with a detailed discussion of all the results (chapter 16). Here the starting points are the aims as specified initially at the end of chapter 2. Set against these, the results are evaluated to determine whether the aims have been achieved. Chapter 17 lists the areas not covered at all or only partly covered in the present thesis. Areas of possible future or related research that are seen to go beyond the scope of this thesis are suggested. The main conclusions to be drawn from the research are presented in a short final chapter 18 for quick reference.

All chapters within parts I to V are structured similarly, in that they begin with a diagram similar to the above indicating placement of the chapter within the thesis. Where appropriate final sections within chapters contain short summaries. A comprehensive list of all references is given in the chapter 19 just before the appendix. The appendix contains detailed results not suitable for inclusion in the main results chapters 13, 14 and 15.

I: INTRODUCTION

2. Introduction and Aim of Thesis

2.1 Discriminant analysis and classification

2.2 Typical examples of discriminant analysis

2.3 Outline of central issues and aim of research

II: REVIEW

III: METHOD

IV: RESULTS

V: DISCUSSION

"Can you please tell me where I ought to go?" asked Alice. "Well it all depends on where you want to get to." replied the cheshire cat.

Chapter 2 - Introduction

This chapter introduces the main issues of discriminant analysis when applied to discrete data. It has been kept non technical, but, nevertheless, intends to introduce the main issues of discriminant analysis when applied to discrete data. Initially the chapter gives a cursory overview of classification problems comparing and contrasting the fields of discriminant analysis, cluster analysis and pattern recognition in the way of a general introduction to the subject. Some essential basic notation is introduced where required. Formal definitions follow later in the review and method chapters. The introduction is practically void of any references. The main body of references with particular emphasis on recent work is given in the literature review in part II. For completeness the review also includes references to areas not directly covered in the thesis such as some of the indirect procedures and selected features of nonparametric density estimation. The central discriminant problem is illustrated by way of examples using artificial data. Consideration of these examples leads to the core issues to be addressed in the research. These are conveniently phrased in terms of a check-list of questions. The introduction ends with a specification of the research aim.

2.1 Discriminant analysis and classification

An interest in *classification* permeates many scientific studies and also arises in the contexts of many applications. For problems such as speech and speaker recognition in acoustics, numerical taxonomy in biology, the detection of diseases by symptoms in health sciences,

the classification of articles in archaeology, the identification of market segments in market research - the central interest is in classifying *objects*, *subjects*, or *entities* of some kind. When the classification is based on measurements of a set of characteristics or variables, statistical techniques are available to aid the classification process.

One may distinguish two broad categories of classification problems. In the first, one has data from known or prespecified groups as well as observations from entities whose group membership, in terms of the known groups, is initially unknown and must be determined through the analysis of the data. For instance, one may have several repeated utterances of a specific word by different persons, and acoustic parameters extracted from each utterance labelled by the particular speaker would constitute the known replicate representations (also called training samples). In such a situation, if some additional utterances of the same word become available but one does not know from which person these utterances arose, one may need to make such an assignment statistically. This is an example of the so called speaker recognition problem where classification is with respect to the known speakers (groups). In the pattern recognition literature reviewed by Duda and Hart (1973) this type of classification problem is referred to as *supervised pattern recognition* or *learning with a teacher*. In statistical terminology it falls under the heading of *discriminant analysis*.

On the other hand there are classification problems where the groups themselves are *a priori* unknown. The primary purpose of the analysis is to determine the groupings from the data so that entities within the same group are in some sense more similar or homogeneous than those that belong to different groups. Many problems of numerical taxonomy, as well as market segments that are determined on the basis of demographics and psychographic profiles of people, provide examples of this second type of classification problem

where the groups are data dependent and not prespecified. This type of classification problem is referred to as *unsupervised pattern recognition* or *learning without a teacher*, and in statistical terminology falls under the heading of *cluster analysis* or *latent structure analysis* rather than discrimination.

Although discriminant analysis and cluster analysis constitute a useful dichotomy of classification approaches, there are many real life problems that combine the features of both situations. One might have some preliminary or imprecise idea of the groups from which the data arise but may seek corroborating evidence of the meaningfulness of the prespecified groups in certain problems. Some combination of the tools from the two types, or perhaps entirely different and as yet unavailable tools, may be appropriate for these situations.

The discriminant analysis situation is characterised by the following: one has two types of multivariate observations - the first, called *training samples*, are those whose group identity (i.e. membership in a specific one of say g given groups is known *a priori*), and the second type, referred to as *test samples*, consists of observations for which such *a priori* information is not available and which have to be assigned to one of the g groups.

The variables constituting the multivariate observations and the "groups" involved will depend on the particular application. For instance, in anthropometry, the variables might be different measurements on fossils and the groups might be a known taxonomy of the fossils (e.g. different races or different stages of evolution). In a medical application, the variables could be the results of various clinical tests and the groups could be collections of patients known to have different diseases. In an acoustical application, the variables might be a set of acoustical parameters extracted from the utterance of a specific word by an individual whereas the groups are repeated utterances

of the same word by different individuals. In each of these cases, there are observations whose group identity is known (the *training samples*) but there will also be some observations whose classification is unknown (e.g. a fossil whose race group is unknown, a patient whose disease category is unknown or an utterance whose source speaker is unknown).

In thinking of the more numerically oriented methods of discriminant analysis, it is useful to distinguish two stages of the analysis, although not all of the available statistical methods either make such a distinction or are equally useful for the two stages. The first stage, concerned solely with the training samples, is to find a representation of these observations so as, in some sense, to clearly separate the g groups. The resulting representation, usually a spatial one, is often called the *discriminant space*. Such a representation when presented graphically has major descriptive and diagnostic value in analysing data.

The second stage of a discriminant analysis is concerned with assigning the test samples (i.e. those observations whose group identity is initially unknown) to one of the g specified groups. At this stage, the focus is on *correct classification*. Some measure of correct classification, using the training samples and not the test samples, is often used to evaluate the performance of discriminant analysis methods. An important scientific consideration, sometimes not emphasised adequately in the statistics literature on discriminant analysis, is that in the real world it may turn out that an item whose classification is unknown may not belong to any of the prespecified groups but indeed be a member of an entirely different or hitherto unknown group (see Rao, 1960, 1962; Andrews, 1972).

Statistical considerations in discriminant analysis have to do with distributional assumptions concerning the observations, measures of separation among groups,

algorithms for carrying out both stages of the discriminant analysis and the study of the properties of proposed algorithms. Historically Fisher (1936) was the first to propose a procedure for the two-group ($g=2$) case based on maximising the separation between the groups in the spirit of analysis of variance. This procedure is equivalent to the likelihood ratio procedure that arises if one assumes multivariate normality with a common covariance matrix for the observations from both groups. Initial extensions were concerned with multiple groups and with heterogeneous covariance matrices across more than three groups, but still retained the assumption of multivariate normality. These normality based methods are the ones most widely used in practice. Provided the measured variables are not constrained to take on only a few distinct values, as in the case of binary variables, transformations of them might enhance their normality and enable the more sensible use of the normality based procedures. An example of this is the frequently skewed (Poisson) distribution of counts that may be "normalised" by either a logarithmic or a square root transformation. There are real situations involving variables, such as binary or categorical ones, that are not sensibly transformed. Distribution free and nonparametric methods, which move away from the normality assumption, have been developed relatively recently to handle such data (see Lachenbruch, 1975; Hand, 1981).

2.2 Typical examples of discriminant analysis

To illustrate the discriminant situation consider first the simplified case of objects belonging to either of two populations from whom continuous measurements on two variables have been obtained. An often used model is to assume that the joint distribution of the independent variables is normal. This is the case in the first example (figure 2.2-1). The data were generated artificially by independent sampling from two bivariate normal

distributions with different means and identical variances and covariances.

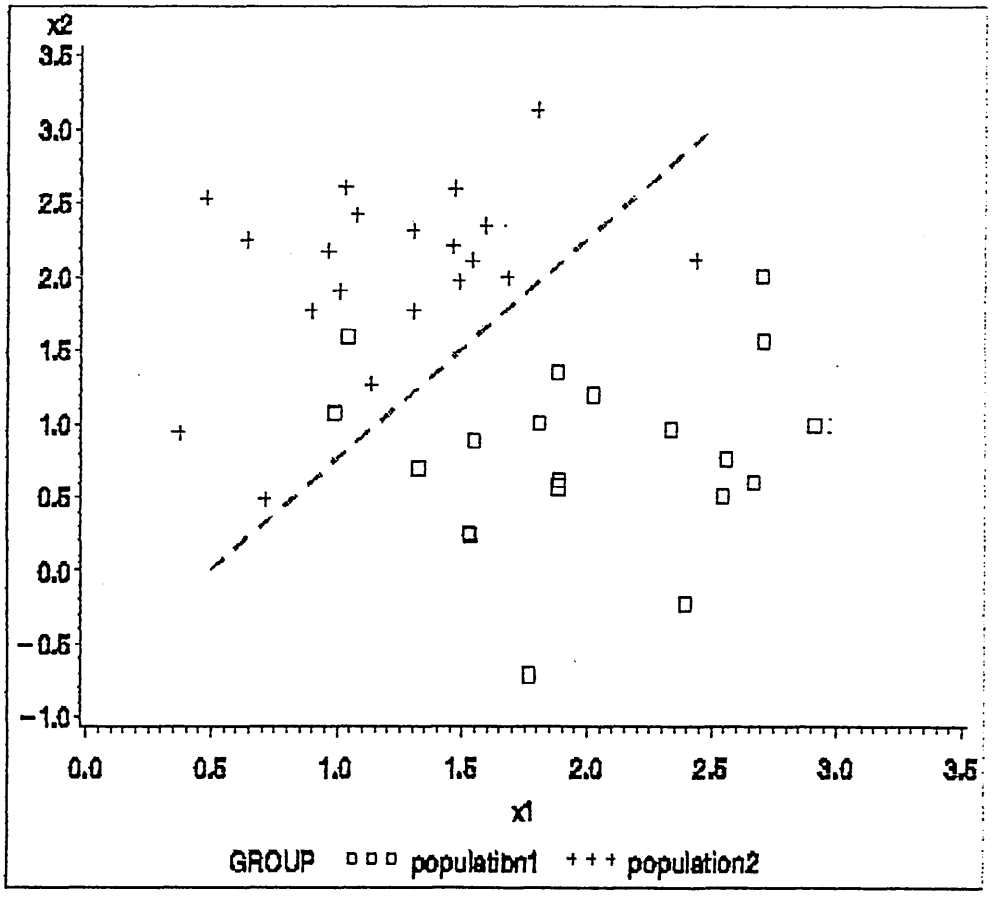


Figure 2.2-1: Continuous data discrimination

Figure 2.2-1 shows simulated bivariate normally distributed data in two populations. The obvious discriminant rule is: *"Allocate any new object whose population membership is unknown to population 2 if its coordinates for X_1 and X_2 lie in the region above the broken (discriminant) line."*

The two populations are clearly visible and a best separation boundary can be specified in terms of a simple straight line in the plane defined by the two independent variables X_1 and X_2. Frequently in situations such as this a straight line provides optimal separation. Occasionally further improvement may be achieved by using curved lines generally described mathematically in terms of polynomial functions. A well known such example is the so called

quadratic discriminant function. The variances are small compared with the differences in means for the variables X_1 and X_2 and so the resulting scatters fall into almost distinct groups. An arbitrarily chosen straight line enables almost complete separation with only *3* observations incorrectly allocated to the other group. Misallocations are points belonging to population *1* yet falling in the region assumed for population *2* and vice versa. Figure 2.2-2 shows the data from figure 2.2-1 after a *discretising* transformation of X_1 and X_2.

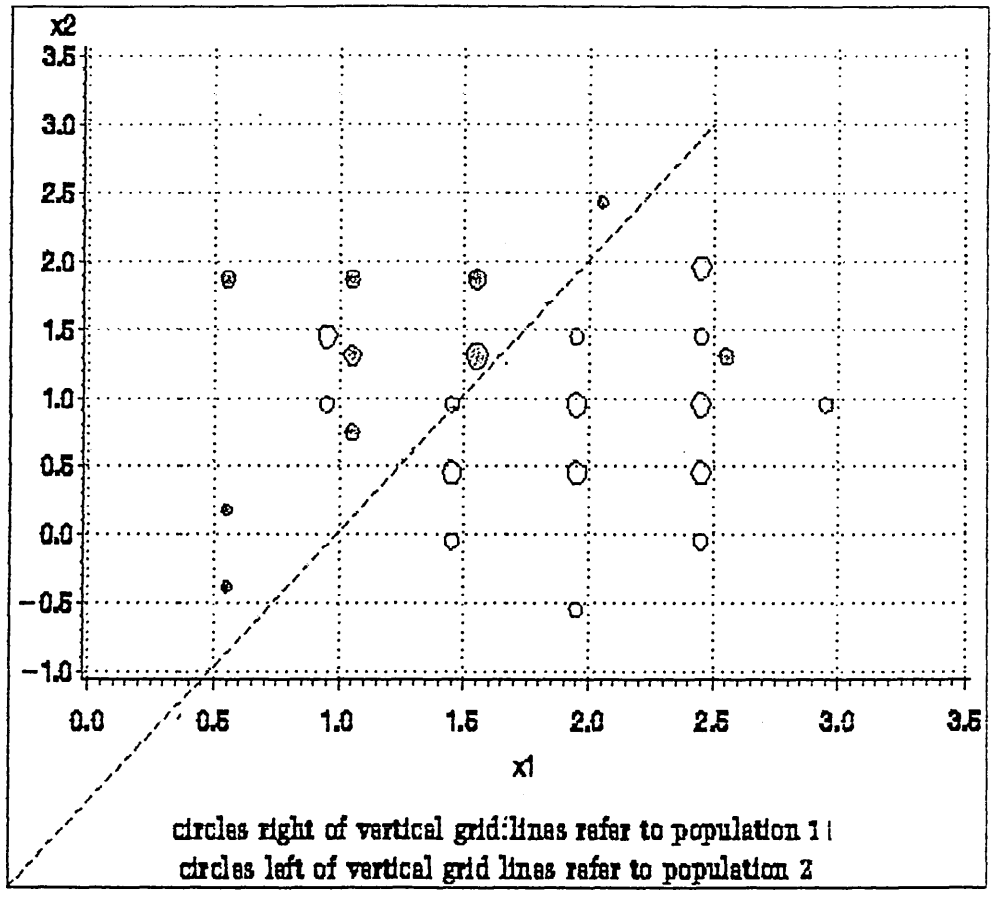


Figure 2.2-2: Discrimination with discrete data

The transition from continuous to discrete data may lead to a loss of information. For the data above no straight line can be found that separates both populations as well as in figure 2.2-1. Medical and survey data frequently take this discrete shape. Small circles resemble *1* observation, large ones *2* observations. To enable distinction observations

from population *1* are displaced to the right of grid intersections and observations belonging to population *2* to the left. The broken discriminant line has been retained from figure 2.2-1. Discretising means that the formerly continuous variables X_1 and X_2 are now only allowed to take on certain discrete values. Here this is achieved by rounding X_1 and X_2 to the nearest *0.5* - other values are of course also possible. As a result of discretisation observations from both populations may coincide. To enable identification of observations from both populations the circles corresponding to population *1* have been slightly displaced to the right of grid intersections and those belonging to population *2* to the left. The area of the circles is proportional to the number of observations with coincident coordinates. The dotted line has been retained from figure 2.2-1. The misallocation errors appear larger with the discretised data. Even other lines would not substantially improve separation. The transition from figure 2.2-1 to figure 2.2-2 demonstrates the potential loss of information when dealing with discrete as opposed to continuous data. This is mainly due to the observations from population *2* with coordinates (*1.0, 1.5*), (*1.0, 1.0*) and (*1.5, 1.0*) lying well inside the area dominated by population *1*. A common example of such information reduction is the transformation of continuous assessment scores out of a hundred to *k*-category examination grades.

Typically a large number of questions in medical research revolve around detection of a suspected disease given observed clinical or laboratory data. Common multivariate approaches to the problem of predicting among several outcomes are multiple regression and discriminant analysis. Statisticians are notorious for pointing out that the assumptions for application of classical discriminant analysis are rarely met under such circumstances, particularly in the fields of medical, social and psychological research. In early years when alternative tools were not as established this used to be a serious issue. However, with the advent of other sophisticated

nonparametric procedures such as recursive partitioning and neural networks in recent years - made possible largely by the ever increasing improvement of hardware and software capabilities - the resistance by some statisticians towards applications of discriminant analysis for discrete data needs reconsidering. With readily available software for nonparametric discriminant analysis violation of the stringent normality assumptions of classical discriminant analysis can be more easily avoided.

The use of classical linear discriminant analysis can be demonstrated by means of literature search. This was done for a sample of recent articles published in the medical field. The *MEDLINE* literature search program provided by the *SILVER PLATTER* company to large university libraries gives access to all papers referenced by the *INDEX MEDICUS* and as such has become a standard reference in medicine. This index was inspected for publications in the years 1989, 1991 and 1993 using the search expression ("discriminant analysis" and either "linear" or "stepwise"). Of the approximately *1000* references on "discriminant analysis" *230* included references to either "linear" or "stepwise". In *148* of these, further information on the type of analysis conducted could be gleaned from the abstracts. On average the number of observations per analysis was *416* with a mode around *100*. Generally the number of variables considered initially (*9.3* on average) was larger than the number of variables finally considered to be related to the outcome event (*5.2* on average). From the information given in the abstracts the type of variable was classified on a dimension ranging from continuous through ordinal and unordered categorical or nominal to dichotomous as shown in table 2.2-1.

type of variable	typical examples
continuous	temperature reaction times weight, height percentages blood cell counts total scores (questionnaire)
ordinal	age examination grades n-point scale (questionnaire)
oininal	blood group colour, race
binary	sex yes/no

Table 2.2-1: The data type dimension

The second column of the table gives typical examples for each type of variable. The boundaries between variable types are deliberately not sharply defined. For instance while temperature clearly is a continuously measured quantity examination grades could be viewed as either continuous or ordinal depending on the underlying data model. When grade is expressed as a percentage it is reasonable to view grade as a continuous variable. However, when grades are expressed as codes ranging from *A* through to *E* they will generally be seen as ordinal. Blood group or sex on the other hand are clearly identified as being of unordered categorical or nominal and dichotomous or binary respectively.

The above classification was used to classify the *148* abstracts according to the types of variable used in the

reported discriminant analyses. It is not uncommon to consider mixtures of variables and so the abstracts were classified by the "highest" variable type with continuous seen as "high". An abstract reporting an analysis based on *6* dichotomous variables and *1* continuous variable is still classed as continuous. Of all abstracts *35* percent referred to analyses based on variables of at most ordinal nature. Even if only variables of at most nominal nature are considered then *9* percent still remain. These findings clearly indicate the continued widespread use of the linear discriminant function under violation of the normality assumptions. The rate of *9* percent even rises to *14* percent when only those variables finally entered into the discriminant function are considered. Of the *148* abstracts *86* percent referred to discrimination between *2* groups, *10* percent to discrimination between *3* groups and only *4* percent to discrimination between *4* and more groups. *97* percent of the abstracts indicated that prediction was the chief purpose of the discriminant while only *3* percent of the reported studies were concerned with analysing the underlying structure of the data. In *84* percent no crossvalidation of estimates was carried out. *Leaving-one-out* crossvalidation or *separate training and test sets* were used in only *16* percent of reported studies.

The following example may help to illustrate the ubiquity of discrete data in medical research. A frequent question in obstetrics but also of general relevance to maternal child health services concerns the management of premature delivery. The causes are still not fully understood. No one single cause has been identified that accounts for all premature births which suggests strongly multifactorial effects. Those factors most likely to be of relevance such as demographic characteristics and psychosocial effects like social class, smoking and stress due to one-parent-family rearing but also past obstetric history, are often inherently difficult to measure precisely and are therefore often dichotomised, i.e. one is dealing with discrete data. It makes more sense to classify a smoker as such if she

smokes heavily than to attempt fine gradings of the number of cigarettes smoked per day as this will be confounded with reporting bias. Another typical example of discrete data is given by Cole et al (1991) who derived a scoring system to quantify illness in babies under *6* months of age using logistic regression on four dichotomous variables.

It was already pointed out for the data in figure 2.2-1 that occasionally curvilinear separations may be required to improve separation in two dimensions. The hypothetical data plotted in figure 2.2-3 shows such a scenario.

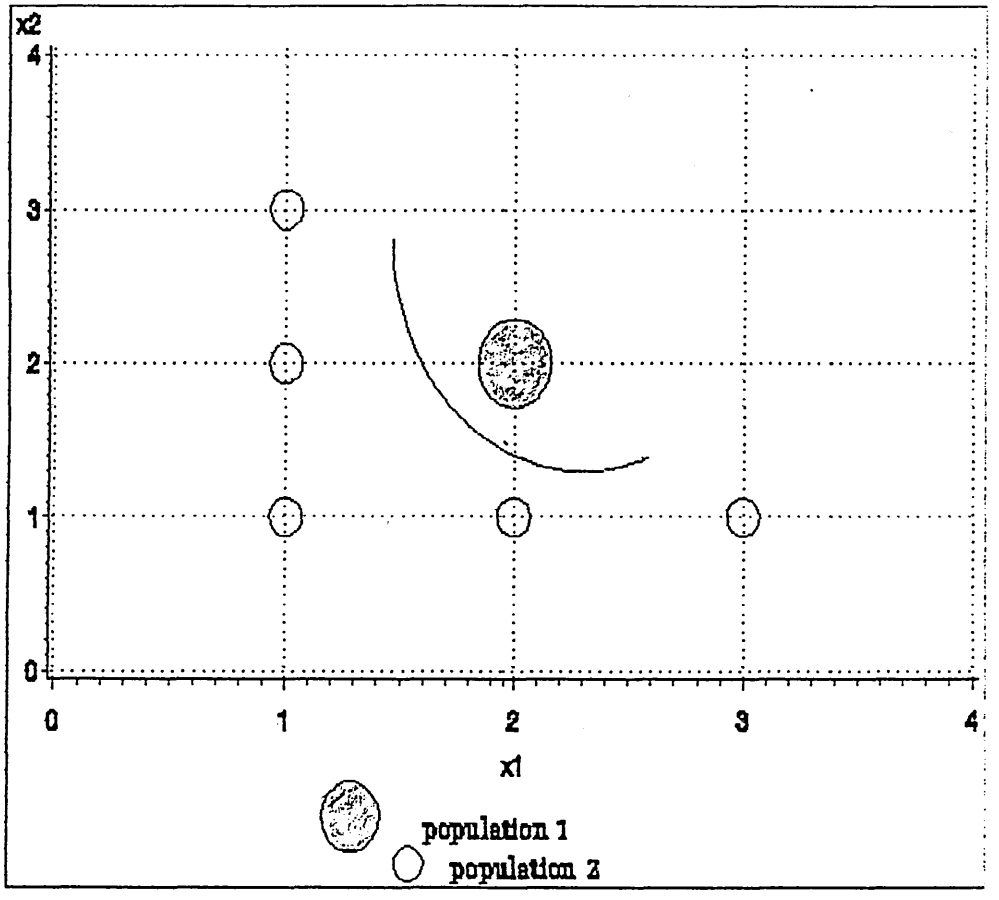


Figure 2.2-3: Curvilinear separation

Figure 2.2-3 shows the obvious discriminant rule for hypothetical bivariate discrete data in two populations as: "*Allocate any new object whose population membership is unknown to population 2 if its coordinates for* X_1 *and* X_2 *lie in the region above the curved (discriminant) line - or equivalently, if variables* X_1 *and* X_2 *jointly exceed about*

1.5." The arc represents perfect separation between both populations. A curve need not be the only solution, however. To see this assume that the discrete levels of X_1 and X_2 refer to *low, medium* and *high*. Next assume that the proximity between the levels *medium* and *high* is greater than between *low* and *medium*. When the data are replotted taking this into account the consequence is that all points except the one at the lower left corner drift out away from the coordinates *(0,0)*. Figure 2.2-4 shows how now again a straight line - mathematically the simpler model - is sufficient for complete separation.

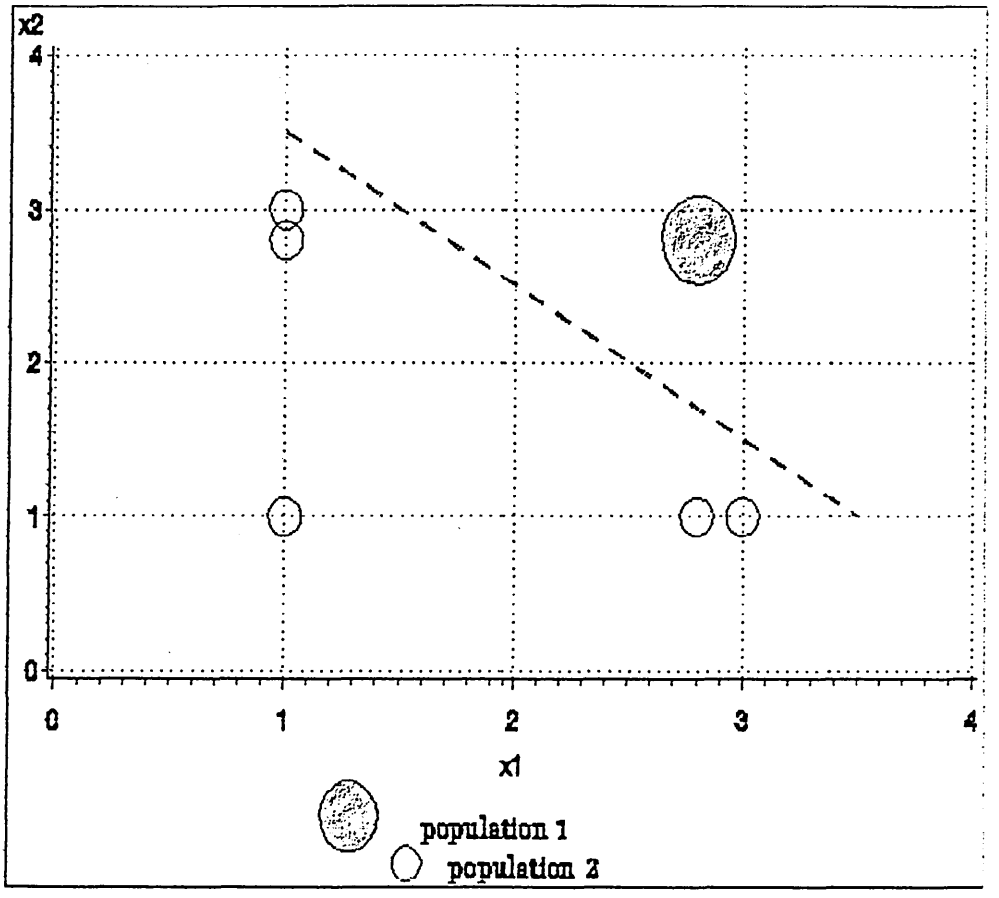


Figure 2.2-4: The effect of scaling

Figure 2.2-4 shows essentially the same data as in figure 2.2-3 yet with a shift of scale values of *2.0* for X_1 and X_2 to about *2.8*. All but one point move away from the coordinate *(0,0)*. Now again the mathematically simpler straight line is sufficient for perfect separation. The obvious discriminant rule may now be expressed more simply

in terms of a straight line: "*Allocate any new object whose population membership is unknown to population 2 if its coordinates for X_1 and X_2 lie in the region above the straight (discriminant) line.*"

The scale shift example has shown how it may lead to a more *parsimonious* solution of the discriminant problem. The *law of parsimony* states that one should generally opt for simpler explanations when there is no obvious evidence pointing to the more complex solution. The next example demonstrates parsimony but this time in relation to sampling. Assume again a bivariate distribution for *2* populations with some overlap such that observations with high values on variable X_2 are predominantly from population *1* and observations with low X_2 values are predominantly from population *2*. Assume further that a sample of an equal number of observations from both populations exists (figure 2.2-5).

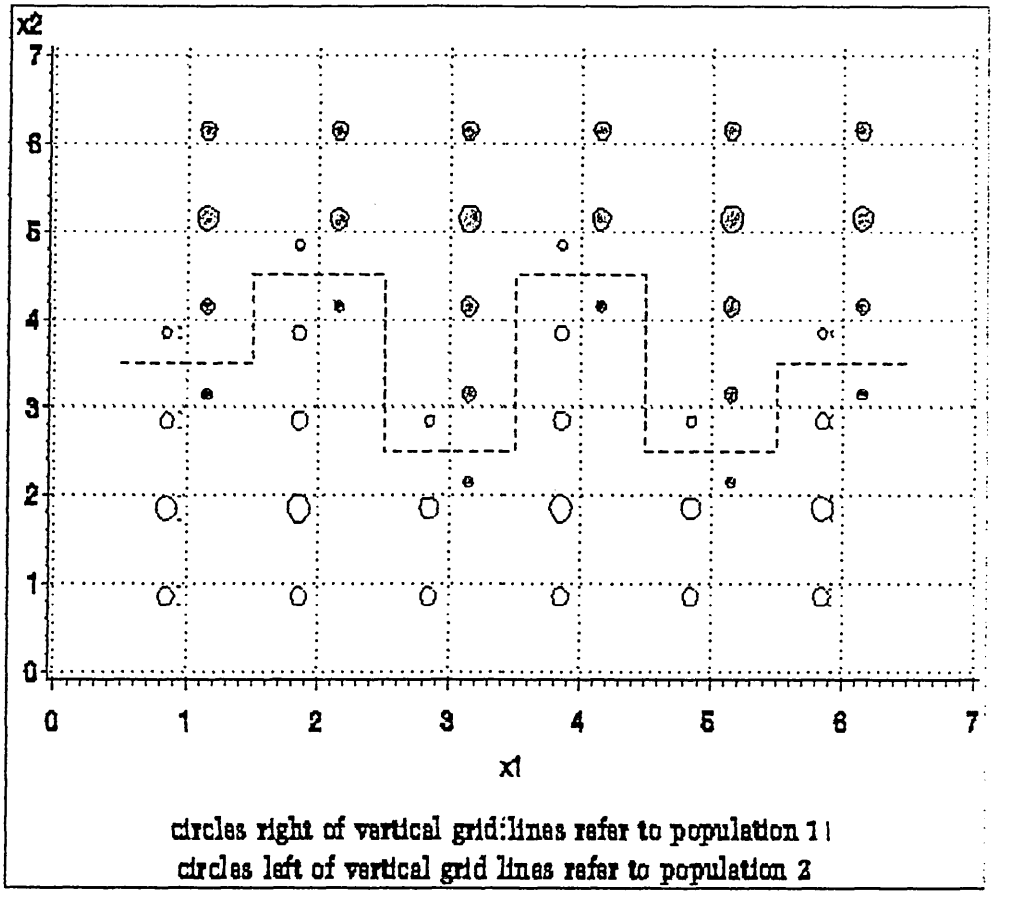


Figure 2.2-5: Optimal fitting of discriminant

Figure 2.2-5 shows an hypothetical sample of bivariate discrete data. The oscillating dashed line has been drawn such that separation leads to least misallocations. Given only this sample an immediate discriminant rule may be: "*Allocate any new object whose population membership is unknown to population 1 if its coordinates for* X_1 *and* X_2 *lie in the region above the broken line.*" Small circles resemble *1* observation, medium ones *3* observations and large ones *5* observations. To enable distinction observations from population *1* are displaced to the upper right of grid intersections and observations belonging to population *2* to the lower left. Separation based on this line in figure 2.2-5 would result in a minimum number of objects from population *1* to be allocated to population *2* and vice versa.

As far as the given sample in figure 2.2-5 is concerned this line represents an optimal solution to the discriminant problem of separating population *1* from population *2*. Next assume that further samples become available (figure 2.2-6) such that the solution depicted in figure 2.2-5 must be considered tentative.

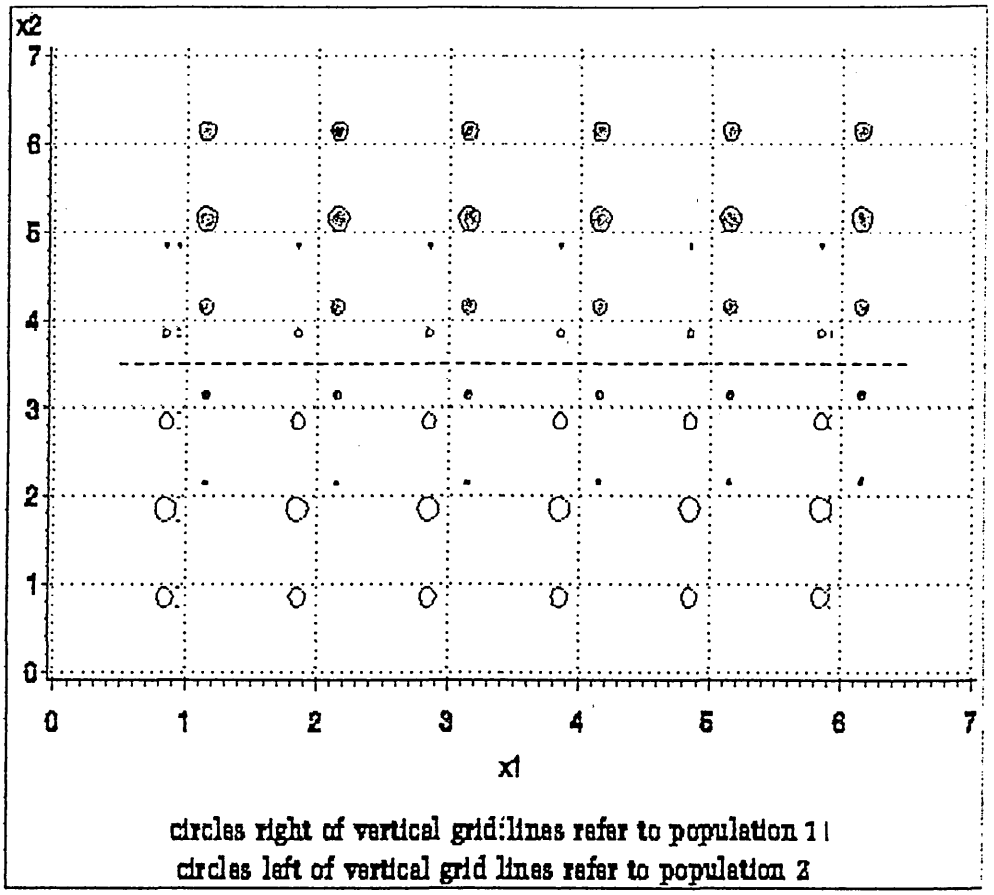


Figure 2.2-6: Effect of sampling further data

Assume that the data shown in figure 2.2-5 comprise one of many samples. Then upon drawing several more such samples the new pattern to evolve may be as shown in figure 2.2-6. Best separation is now achieved by a simple straight division line. Again the discriminant rule is: "*Allocate any new object whose population membership is unknown to population 1 if its coordinates for X_1 and X_2 lie in the region above the broken line.*" However the straight line model of the separation line is simpler. By the law of parsimony this is the favoured solution which might be missed when sampling is inadequate. Due to sampling the apparently oscillating pattern suspected from figure 2.2-5 has vanished and suggests a simple straight line as the more parsimonious solution. In fact the solution here is even simpler because optimal separation is possible using X_2 values alone as the best separation line is parallel to the X_1 axis. This example shows the danger of choosing overspecified models when sampling is insufficient.

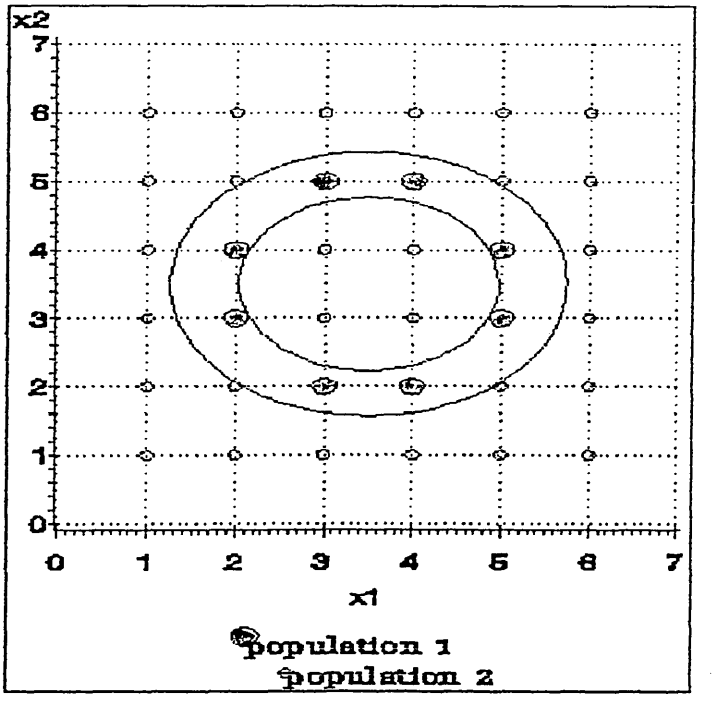


Figure 2.2-7: Pattern recognition

Figure 2.2-7 shows another hypothetical example of bivariate discrete data with observations placed at unit intervals. However, this time the data form distinct patterns. The two populations are again distinguished by different sizes of circles. Observations belonging to population *1* are arranged in a distinct pattern. However, no single dividing line enables complete separation of the two groups. The above example is less common in discriminant analysis and more typical of the type of problem faced in the field of pattern recognition. In the present case the question may have been to detect the symbolic representation of the letter "*O*" in front of a uniform background.

Frequently multivariate data may show some degree of correlation among the "independent" variates. For instance when X_1 is high X_2 will also be high. When the centres of the scatters for both populations are sufficiently far apart it is easy to divide them by a single straight line.

Often however multivariate data may in addition reveal *interactions*. Loosely speaking *interactions* are present when the correlation between variables changes according to the values of one of the variables.

Figure 2.2-8 shows bivariate discrete data with interactive effects between X_1 and X_2.

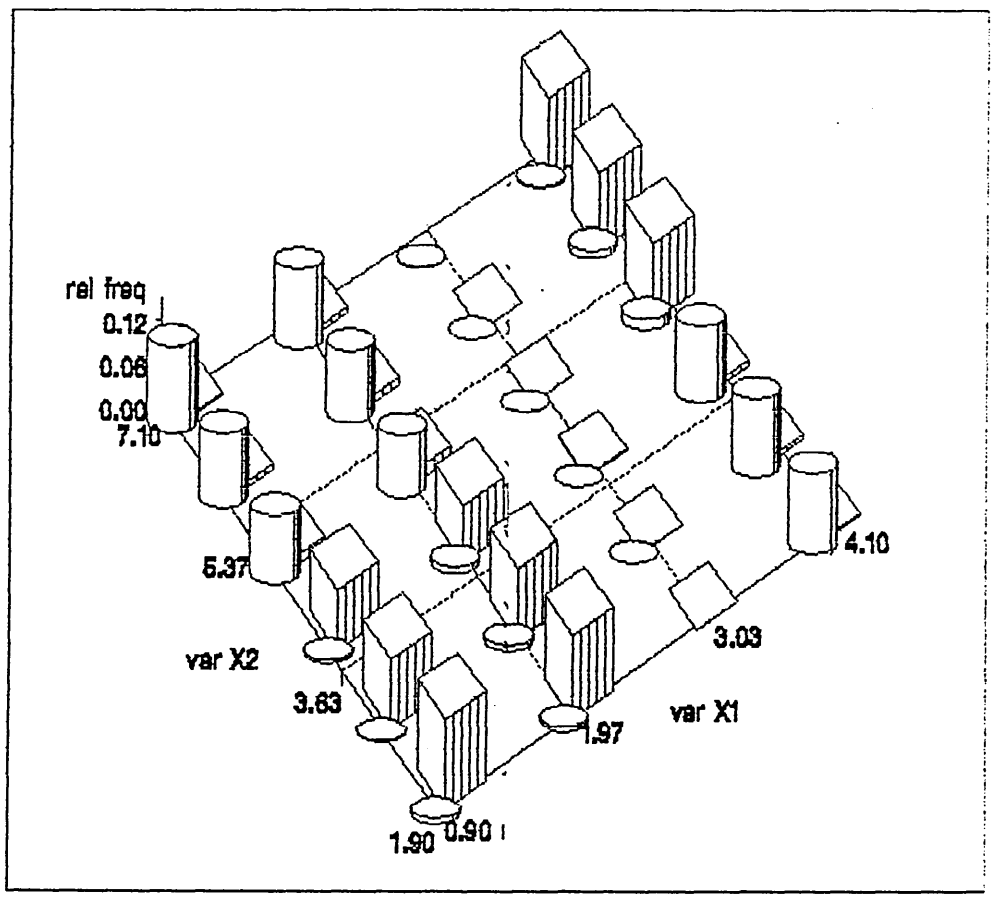


Figure 2.2-8: Bivariate discrete data

Populations *1* and *2* are now spread out, with observations from population *1* occupying the lower left and upper right corners. Clearly no single discriminant line can be found for optimal separation. Instead two lines would be required, as may be seen from figure 2.2-8. Observations from both populations are largely separated into opposite corners of a square. Two discriminant lines are required for optimal separation.
A similar situation is presented in figure 2.2-9.

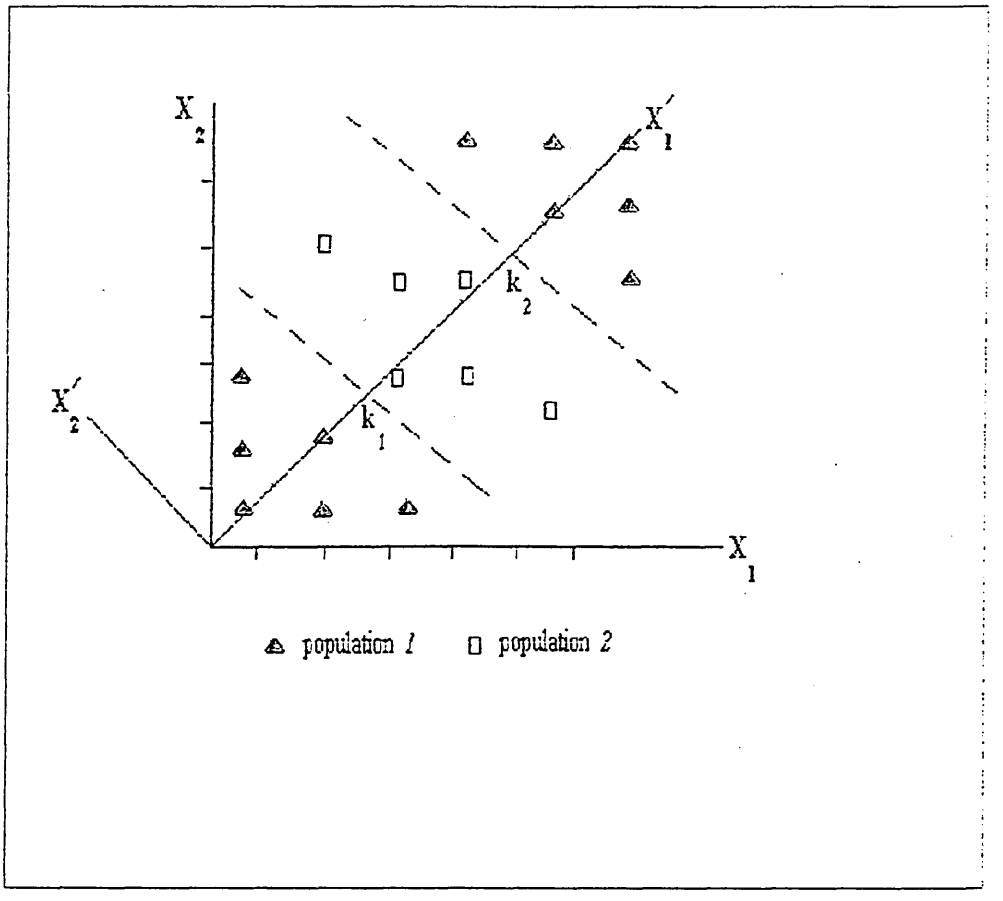


Figure 2.2-9: Bivariate discrete data

Figure 2.2-9 shows bivariate discrete data with observations from population *1*, split into opposite corners of a square. The standard solution in terms of the original coordinates X_1 and X_2 is indicated by the dashed lines. Rotation of the coordinate axes anticlockwise to X_1' and X_2' yields a simple discriminant rule in terms of only X_1': "*allocate to population 2 if* $k_1 \leq X_1' \leq k_2$ *and to population 1 otherwise*", where k_1 and k_2 are suitably chosen cut off points.

Observations belonging to population *1* are split into two lumps placed at opposite corners of a square. Two obvious discrimination lines have been drawn in. The discriminant problem may be simplified however by observing that rotation of the coordinate axes through the two centroids of population *1* enables specification of the discriminant rule in terms of only one variable as follows: "*allocate to*

population 2 if $k_1 \leq X_1' \leq k_2$ *and to* population *1 otherwise*" where X_1' is the rotated original X_1 axis and k_1 and k_2 are suitably chosen cut off points. The example shows how transformation of original variables to more suitable ones may simplify the discriminant rule.

Occasionally the data may exhibit considerable overlap between the populations to be differentiated in the central region yet indicate reasonable separability in the tails of the distributions. An example for 3-level independent variables X_1 and X_2 is constructed as a bubble graph.

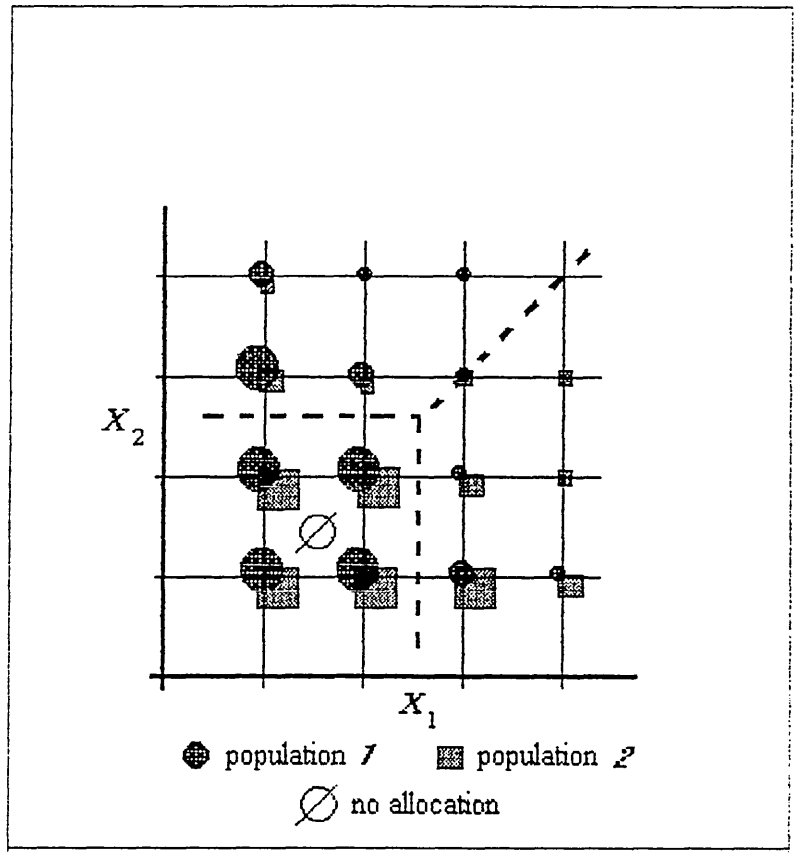


Figure 2.2-10: Bivariate discrete discrimination

Figure 2.2-10 shows bivariate discrete data showing large overlap at coordinates for which X_1 and X_2 are less than 2. Useful discrimination appears possible only outside these points. Corresponding discriminant lines are drawn in. Size of the bubbles indicates number of observations at each

location. It can be seen that outside these *4* coordinates allocation appears quite clear. The discriminant lines drawn in figure 2.2-10 thus exclude the doubtful coordinates and the rule might be "*focus discrimination only on the remainder of observations where* $X_1 > 2$ *or* $X_2 > 2$".

The above examples presented in figures 2.2-1 to 2.2-10 illustrate the spectrum of problems addressed in the thesis. These may be summarised in the form of typical questions pertaining to the discriminant situation. As these questions are central to the application of discriminant analysis, especially in the case of discrete data, any formally structured approach to procedure selection should provide answers to all of them.

(a) *Is there reliable distributional information available to allow selection of a specific customised parametric procedure ?* The functional form of the data's distribution (Poisson, normal, binomial, ...) must be known. If the values of the respective parameters (λ, μ & σ^2, n & p) are not available then it should be possible to compute reliable estimates from the given sample.

(b) *Is the given data sample large enough to provide reliable estimates of distributional parameters in the cases where prior information points to highly complex data models and could a simpler model be just as efficient ?* The knowledge that the underlying data model requires a large number of parameters to be estimated (for instance in the case of the full multinomial model where the number of parameters equals the number of cells) will be of limited use if one is faced with a small dataset. Stable parameter estimates require large data samples.

(c) *How does one proceed when one is not sure of the available distributional information ?* This question opens up the field of nonparametric density estimation and of alternative procedures such as recursive partitioning, projection pursuit and neural networks.

(d) *Is complete separation of the populations possible ?* There may be structurally inherent reasons that rule out a certain degree of unavoidable overlap a priori no matter what discriminant procedure is selected. It is not uncommon for certain individuals to be indistinguishable in terms of the available information. For instance diseased and healthy subjects in a medical study may exhibit identical symptoms. Certain newer discriminant procedures (such as some of the simpler neural networks) will only function if complete separability can be guaranteed (see chapter 6).

(e) *How does one deal with indistinguishable objects ?* They may be allocated by chance to either of the populations, they may all be considered as misallocations or they may be excluded altogether.

(f) *How does one deal with marginal allocations ?* The attribute *indistinguishable* may be relaxed to mean *similar*. Similar objects from different parent populations differing only slightly in their characteristics cast doubt on the certainty of a derived allocation. Specifying that some difference measure between objects should exceed a suitably defined minimum value may help to improve overall reliability of a discriminant procedure.

2.3 Outline of central issues and aim of research

Previous work in the field of discriminant analysis was carried out in the context of an MSc project at Sheffield City Polytechnic (Lack, 1987). That work involved finding suitable predictors for the incidence of still births using data collected in routine surveys of maternal child health monitoring schemes. The discriminant procedures used then were to some extent arbitrarily chosen. At the end of the initial construction of predictors of stillbirth it was felt that there was clearly a need for further research as there appeared to be few guides to choosing the right procedure available in the literature. Early work on the present PhD thesis confirmed this view.

Today a need for further work is evident:

(a) in the field of discrete data with a limited number of discrete states, and

(b) in particular concerning the provision of guides to optimal procedure selection.

To achieve this the following steps were undertaken:

(a) literature review,

(b) construction of performance criteria suitable for discrete data, and

(c) construction of a general (and formal) guide to procedure selection suitable for discrete data involving dichotomous, nominal and ordinal variables.

The discriminant analysis literature is abundant with new procedures - recently especially in the nonparametric and mixed data fields - yet gives relatively little guidance on the choice *among* procedures. While this may not be a drawback with most large datasets, especially when the data

are normally distributed, this does present serious problems in the cases of small datasets with discrete data structures. Typically this is the case in many fields of research, especially in medicine as was demonstrated in section 2 of chapter 2.

The work presented below addresses this problem and attempts to fill this gap by suggesting a variety of criteria that will help to guide procedure selection in some *optimal* sense. The emphasis will be on two areas of research identified as scarcely covered in the discriminant analysis literature; the use of information in the distribution of posterior probabilities across discrete states of the data and the employment of classification thresholds. It is expected that both these tools will provide more help in choice of a suitable discriminant procedure.

In several cases direct extensions of the developed ideas to continuous or mixed datasets or to other procedures are possible and will therefore only be indicated but not pursued in detail. Because of the central role played by the distribution of posteriors the work will focus largely on, but not exclusively, so called direct procedures (see chapter 4).

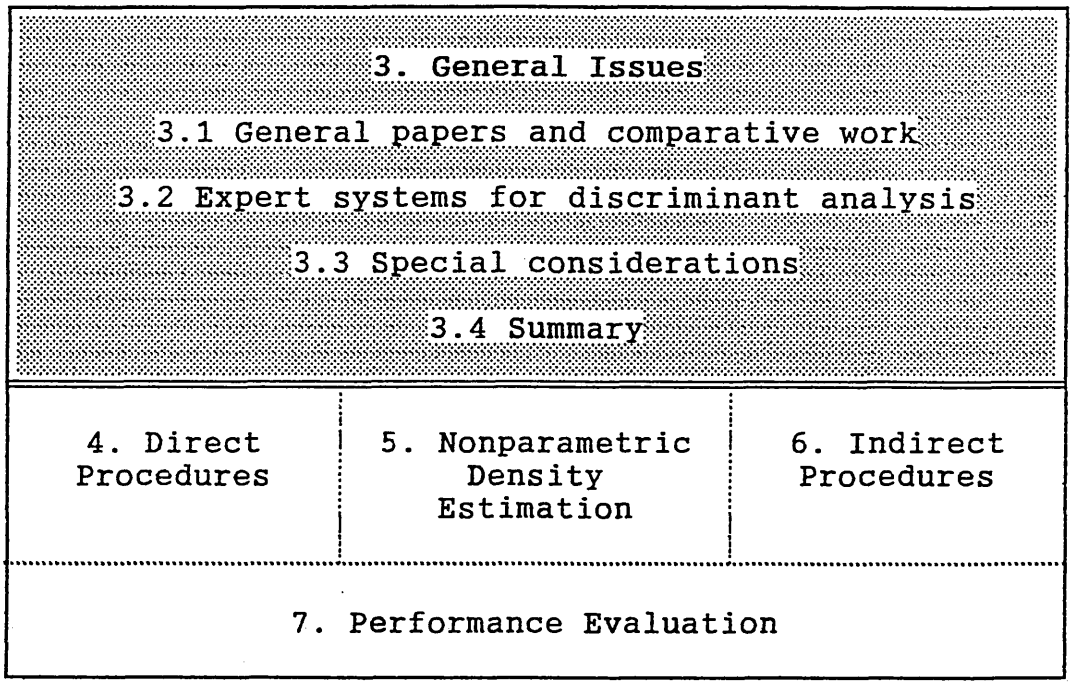
The common likelihood ratio approach operates on a qualitative assessment of the relative size of the posteriors. If the likelihood ratio exceeds a *constant* value - unity in the case of discrimination between two populations - sufficient justification for an allocation is inferred. So allocation is made irrespective of the actual size of the posterior probabilities. As these may vary due to sampling effects, the notion of *thresholded allocation* is introduced. The basic idea is that for situations in which comparatively poor separation is inherent in the data a differential consideration of the *absolute magnitude of the posteriors* may enhance the quality of performance assessment while with relatively good separation of the

populations the commonly applied likelihood ratio approach would be expected to function satisfactorily.

While it may at first appear natural to expect a *cookbook of procedures*, this is not the chief aim. Rather the description and outline of the basic methodology is seen to constitute the core of the presented work. This focus is motivated by the fact that textbooks or journal papers on discriminant analysis applications to discrete data generally highlight selected features of the discriminant problem. An embracing treatment of the entire range of problems addressed by density estimation, sampling issues, reliability of allocations, the technical handling of discrete data in terms of state matrices and last but not least the role of model assumptions underlying the data does not exist to the knowledge of the author.

I: INTRODUCTION

II: REVIEW



III: METHOD

IV: RESULTS

V: DISCUSSION

Chapter 3 - General issues

This chapter is divided into a section 3.1 reviewing general papers and comparative work, a section 3.2 on expert systems for discriminant analysis and a section 3.3 on special considerations regarding transformation of variables, selection of variables and missing data estimation. The emphasis will generally be on applicability to discrete data. This restriction is imposed because of a lack of guides to selection of procedure in the discrete data field. Although many real data situations involve mixed data types (continuous, ordinal, nominal and dichotomous) frequently the raw data can be reduced to lower scales of measurement without substantial loss of information. Typically, studies reporting (discriminant) analyses of mixed data types conclude that the original scales of the variables can be *discretised* - i.e. a continuous variable is transformed to an ordinal one, an ordinal variable is transformed to a dichotomous one.

3.1 General papers and comparative work

Data types may be arranged along a *dimension of discreteness* ranging from continuous at one end of the spectrum to dichotomous or binary at the other end. Table 3.1-1 shows classes of predictor variable used in discriminant analysis applications with examples of typical usage. Continuous variables are included - for the sake of completeness. The response variable is frequently dichotomous, otherwise nominal and occasionally ordinal.

Type of variable	Example		Presentation of data
dichotomous	answers	yes, no	usually counts arranged in a contingency table - the general case for discrete data
	factors	present, absent	
trinomial	political	socialist, liberal, democrat	
multinomial	blood group	A, B, AB, O	
ordinal	levels	low, medium, high	generally as individual observations ("raw" data) but frequently, especially when number of discrete states is not too large, also in shape of contingency table of counts
	counts	0, 1, 2	
	rankings	1st 2nd 3rd	
continuous	temperature	in deg. C.	as individual observations
	weight	in kg.	
	proportions	in %	

Table 3.1-1: Data types in discriminant analysis

Corresponding to this *dimension of discreteness* are respective data models ranging from *product binomial*[2] and *product Poisson* (in the case of multidimensional contingency tables) through *full multinomial* to *multivariate normal* (in the case of continuous normal distributions). Because of the (still present) popularity of the linear discriminant function it is common practice to achieve the required normality assumption of the independent variates by prior *normalisation* of the data using appropriate transformations such as $\log_e(x)$,

[2] Terminology of Bishop, Fienberg and Holland (1975)

$\sinh^{-1}(x)$, $\sqrt{x}$ or rank(x). Clearly this approach has its limitations as is evident when faced with the task of normalising dichotomous data. The problem here is knowing when a transform will reliably produce a sufficiently normal variable and when a different, more appropriate, data model is required.

Unlike other statistical techniques discriminant analysis draws simultaneously on a variety of different statistical tools. Five major areas may be identified:

(1) model assumptions have to be made about the underlying data distributions,

(2) based on these, the generally unknown densities have to be estimated,

(3) suitable measures of distance between populations are required for measuring separation,

(4) performance is assessed by estimation of misallocation errors, and

(5) these are frequently based on posterior probability functions of population membership.

The literature on discriminant analysis may be divided into these five categories in terms of the major emphasis in published articles. The greater part of published work addresses topics (2) density estimation and (4) error rate estimation. Major papers on general issues in discrete discriminant analysis are shown in figure 3.1-1. A discussion of indirect techniques (see chapter 4) appears relatively late in the literature. Selection guides to discriminant analysis are rare. Selection of variables, discrimination between three or more groups, ordered categories, missing data estimation and mixed distributions are excluded.

density estimation	Fix & Hodges	1951
	Martin & Bradley	1972
	Aitchison & Aitken	1976
	Ott & Kronmal	1976
	Titterington	1980
	Titterington et al	1981
	Hall	1981a
	Hall	1981b
	Butler & Kronmal	1985
	Silverman & Jones	1989
error rates & optimality	Cochran & Hopkins	1961
	Hills	1966
	Lachenbruch	1967
	Lachenbruch	1968
	Glick	1972
	Lachenbruch	1975
	Goldstein & Wolf	1977
	Hora & Wilcox	1982
	Efron	1983
	Konig	1988
	Lawoko & McLachlan	1989
	Snappin & Knoke	1989
	Kharin	1990
	Solow	1990
	Williams et al	1990
distance measures	Hills	1967
posterior probability functions	Titterington et al	1981
	Hora & Wilcox	1982
	Goldstein & Dillon	1984
model adequacy	Gilbert	1968
	Christl & Stock	1973
	Victor	1976
selection guides	Pfeiffer	1987
	Wernecke et al	1989
	Patuwo et al	1993
	Lark	1994

Figure 3.1-1: Studies on discrete discriminants

Performance evaluation of a discriminant rule is generally in terms of its *misallocation error*, or equivalently, in terms of its *correct allocation rate*. Corresponding estimators are usually judged in terms of their respective potential for *bias reduction* and *variance reduction*. The methodology involved in discriminant procedures and performance evaluators overlaps to a certain extent. This will become particularly evident when in the context of *crossvalidation* techniques an identical algorithm is used

in one case to produce better density estimates and in another to stabilise a misallocation error estimate.

Stable performance criteria (error rates and reliability measures) require some form of *crossvalidation*. The actual kind of crossvalidation method used will in turn depend on the assumed data model. If no particular assumptions are made then the *bootstrap* based on the observed multinomial frequencies may be appropriate. If, however, an interactive data structure in terms of a contingency table may be inferred then an appropriate generation algorithm, for instance based on the Bahadur (1961a, 1961b) representation of dichotomous data or its extension for discrete data due to Lancaster (1969) might be adopted.

A different view of performance evaluation for discriminant analysis of discrete data is taken by Hills (1967)[3]. Instead of inspecting the allocation results after application of a given discriminant procedure he suggests evaluating a procedure`s performance by judging its qualities prior to application. In a sense his approach is theoretical rather than empirical. He lists four basic properties that distance measures for discrete data should ideally satisfy. Examples of such measures based on computing differences between individual state probabilities are given. Procedures are then defined a priori as efficient if the distance measure employed exhibits these desirable properties. Hills' approach, however, is applicable only to procedures that use differences in individual state probabilities. For example the performance of a procedure based on the generalised population distance (Goldstein and Dillon, 1978) could not be assessed using Hills' distance measures because it is not a function of individual state differences.

A natural source of information in the published literature for a researcher seeking advice on which procedure to apply

[3] see also chapter 6, section 1

to a given dataset will be general papers reporting reviews and comparative analyses. For almost two decades Goldstein and Dillon's (1978) book on discrete discriminant analysis was the most comprehensive piece of work on this subject. It presents the *linear discriminant*, *quadratic discriminant*, the range of *interactive* procedures from the *independent* model through the *Bahadur* procedures to the *full multinomial*, *nearest neighbour* and *logistic* procedures. While these procedures are described and also illustrated with some comparative work, comparatively little advice is given about how to *select* a procedure given a particular dataset. The conclusion reached by the "Panel on discriminant analysis" (Gnanadesikan et al, 1989) appeared to be that there was still a need for suitable guides to selection of optimal procedures. To date only very few references to selection guides for discriminant analysis can be found in the literature. In the extensive searches conducted only two references were found over a period spanning *25* publication years: Wernecke et al (1989) and Lark (1994). This finding has not changed much even with the recent advent of McLachlan's book on discriminant analysis and statistical pattern recognition (1992), which then might have been considered to be a representative and exhaustive source on discriminant analysis for all types of variable, continuous, discrete or mixed. This book contains an updated set of discriminant procedures now including *kernel* procedures and the *CART* procedure. It also discusses the *posterior error rate estimator* which has favourable properties that are useful in comparative analyses. Since 1992 however, the neural network literature has blossomed and from that point of view even McLachlan's book may already be considered outdated.

Major comparative analyses are reported by Gilbert (1968), Moore (1973), Goldstein and Dillon (1978), Titterington et al (1981), Trampisch (1983), Pridmore (1985), Wernecke (1986), Bull and Donner (1987) and Cox and Snell (1989).

3.2 Expert systems for discriminant analysis

So-called expert systems are increasingly popular in many disciplines because they alleviate much of the burden placed on regular users. In medicine for instance, some authors report expert systems in use today, which help identify diseases by matching certain input data with already stored data bases. Oppel (1990), Schewe, Herzer and Krüger (1990), Hadzikadic (1992), Clyma and Lancaster (1993) and Ohmann et al (1995) give recent accounts of some applications of expert systems in medicine. For some of these systems it is claimed that they can *learn*. Some in fact use variants of discriminant analysis for distinguishing between different diseases.

What is fairly commonplace in medicine is still comparatively rare in statistics. To date few expert systems exist to aid the user of statistical software packages in selection of an appropriate tool beyond the standard hierarchical structure of procedures in the sequence of help screens. Publications with titles such as "Which discriminant function should be used?"[4] by Pfeiffer (1987) or "Sample size and class variability in the choice of a method of discriminant analysis" by Lark (1994) are few and far between.

3.3 Special considerations

When presented with new data in the context of a discriminant problem an essential question is whether the given data are optimal in the sense that apart from sample size and representativeness the "right" variables are available. This leads to the topics of *selection of variables* and *transformation of variables*. Respectively, these topics address the questions of which of all

[4] This paper relates to analyses of robustness of the linear discriminant, the logistic and the kernel procedure applied to clinical case-control studies.

available variables are essential and whether the given variables are ideal in their form or whether the discriminant would benefit from individual re-scaling or linear and non-linear combinations of variables.

A further question concerns missing data. Although selection of variables and transformation of variables are related concepts they will be reviewed separately.

3.3.1 Transformation of variables

Transformation of variables prior to applications of discriminant analysis to discrete data may be expected to improve performance. In the case of continuous data transformation is generally used for normalising purposes. Similarly discrete data on observed frequencies of events distributed as Poisson might lend themselves to a logarithm or a square root transformation prior to application of a linear discriminant. What emerges however is a picture where individual skill and experience of users conducting discriminant analyses appears to be crucial. This fact should not however be seen to suggest that transformation of variables is a side issue. The example shown in the introductory chapter 2 (figures 2.2-3 and 2.2-4) illustrated how judicious scaling of explanatory variables may enhance discrimination. Another example bordering on the issue of model selection is that of creating non-linear combinations of the original variables when interactions are suspected. Assume that a main effects model is used for estimating a discriminant function for independent variables X_1 and X_2. If strong interactions are present in the data the discriminant based only on main effects will perform poorly. If however a third variable $X_3 = X_1X_2$ is added as the product of X_1 and X_2 the discriminant will generally be improved. This is perhaps not a classical example of transformation of variables as an appropriate model choice ie, one that models the joint effect of X_1 and X_2, would yield similar results.

Although not treated here in depth, transformation of heavily skewed and obviously non-normal variables prior to performing classical linear discriminant analysis can markedly improve performance. As an example see the work of Friedman (1989) who examined the efficiency of a generalisation of the traditional normal linear, *LDF*, or quadratic, *QDF*, discriminant function. This procedure - which he calls *regularised discriminant analysis*, *RDA* - replaces each normal density used in the traditional classification rule by a Fourier series density estimator which *adjusts* the normal density if the data deviate markedly from normality (eg heavily skewed or multimodal). Friedman (1989) derived the *RDA* procedure in both univariate and multivariate situations. Friedman concludes that if the distributions of the data do not deviate markedly from normality then *RDA* is as efficient as *LDA*. On the other hand, if either of the distributions deviates from normality, then *RDA*, which performs as a semiparametric discriminant procedure, is more efficient than *LDA*.

Transformation of variables, though not unimportant, is treated as a side issue in the present thesis as the emphasis is on developing appropriate tools for choice of an *optimal* procedure from a range of procedures on the basis of given datasets.

3.3.2 Selection of variables

In contrast to transformation, the selection of variables has received greater attention in the literature. This is presumably because selection lends itself more readily to formal approaches. Whilst transformation demands finding the most suitable type of transformation from among an ultimately infinite range, selection of an ideal subset p from q variables demands consideration of at most $\binom{q}{p}$ combinations. Selection of variables becomes an issue when

either the number q of available feature variables is large relative to total sample size or when costs of gathering and processing large sets of variables proves prohibitive. Common choices of criteria for subset selection are misallocation error rates, F-ratios, distance measures and information measures such as the divergence measure of Kullback and Leibler (1951). Studies addressing selection of variables for discrete data include those by Elashoff, Elashoff and Goldmann (1967), Hills (1967), Goldstein and Rabinowitz (1975), Goldstein and Dillon (1977), Goldstein and Dillon (1978, chapter 4), Haerting (1983) and Krusinska and Liebhart (1988, 1989). For a general review of selection of variables see McLachlan (1992, chapter 12). While a variety of techniques for subset selection exists for continuous data this does not hold for discrete data. In the continuous case the stepwise computation of F-ratios to assess variables for exclusion from the discriminant function is standard practice (e.g. Goldstein and Dillon, 1977). In the discrete data situation variable selection guides are so far given only with respect to the recent recursive partitioning algorithms such as *CART* or *FACT* which produce so called *variable importance rankings* based on heterogeneity measures (see Breiman et al, 1984; Loh and Vanichsetakul, 1988; as well as the account in chapter 6).

For a considerable time the *ALLOC1* algorithm developed by Habbema et al (1974) used to be the only technique that was able to handle cases of multiple groups and non-normality. Hermans et al (1982) gave an extension of this algorithm called *ALLOC80*. Habbema et al use the plug-in sample version of the direct rule, where the group-conditional densities are estimated non-parametrically by the kernel method. Their algorithm was the multivariate normal density with a diagonal covariance matrix as the kernel. The smoothing parameter is estimated by the program. A subsequent modification allows the program to use variable kernels to provide better estimates of the group conditional densities. The *ALLOC1* algorithm achieves subset selection in terms of the overall error rate.

Variable selection techniques may be split into two types, direct procedures and stepwise ones. Direct methods, as used by Pipberger et al (1968) and Lack (1987), include entering variables in their order of univariate statistics derived from formal tests (such as the Wilcoxon, χ^2-tests or t-tests), arbitrary selection or direct entry of all variables en bloc. Stepwise procedures may again be split into restricted ones and unrestricted ones, as in Hills (1967), where variables entered at an earlier stage may be deleted again at a later stage. Stepwise procedures may be based on a number of approaches. These include:

(1) selection by minimising node impurity, as implemented in Breiman, Olshen, Friedman and Stone's (1984) classification and regression trees,

(2) selection by maximisation of a distance measure such as Matusita's (1955, 1956) distance or simply an euclidean measure,

(3) selection by minimisation of an estimate of misclassification error such as done by McLachlan (1976) but rarely used in practice, or

(4) selection by minimisation of residual sum of squares.

Habbema, Hermans and van den Broek (1974) give a procedure based on the estimated values of posterior probabilities using robust kernel density estimation methods. Selection of variables, although it may be of major consequence for the performance of any procedure, is excluded from a thorough treatment in this thesis.

3.3.3 Missing data

Estimation of missing values is a general problem in statistics. With respect to discriminant analysis, methods for the handling of missing data may be grouped into estimation methods and other methods. The former includes substitution of grand means, group means or estimation by

regression of all other variables for which information is given onto the variable containing the missing value as done in the *BMDP* statistical package. Titterington et al (1981), for instance, use means substitution and Murray and Titterington (1978) use kernels. Other methods include treating a missing value of a variable as another distinct multinomial state which is very common in discrete discriminant analysis or by use of *surrogate splits* as done in the *CART* procedure[5]. A comparison of *4* different methods in application to the multivariate binary case is given by Titterington (1977). The methods involve (a) setting the smoothing parameter λ to *1/2* in the Aitchison and Aitken (1976) kernel corresponding to the uniform model, (b) maximum likelihood methods, (c) regarding *missing* as an extra category so that binary variables become *3*-level or *ternary* and using a version of the kernel estimator of Aitchison and Aitken (1976) and (d) treating *missing* as a category between *symptom present* and *symptom absent* thus treating the categories as ordered.

In *CART* (Breiman et al, 1984) the missing data algorithm is designed to accomplish two purposes simultaneously: first to make maximum use of the data cases, complete or not, in the tree construction and second to construct a tree that will classify any case dropped into it even if the case has some variable values missing. This differs from the usual missing value procedures in regression or classification, where the covariance matrix is estimated and then used to produce a single prediction equation defined only on complete cases. Suppose that the best split s^* on a node is being found. If there are missing values the best split s_m on $\mathbf{x}_m$ is computed using all cases containing a value of $\mathbf{x}_m$ and then s^* selected as that split s_m^* which maximises $\Delta_i(s_m^*,t)$. In linear combinations the best split is computed using all cases complete in the ordered variables. For a Boolean split, i.e. linear combination of variables, all cases complete in the variables appearing in the

[5] see chapter 6, section 2

Boolean expression are used. If a case has missing values so that s^* is not defined for that case then among all nonmissing variables in that case the one is found, x_m, say, with $\tilde{s}_m$ having the highest measure of predictive association with s^*. The case is then split using $\tilde{s}_m$. This procedure is analogous to replacing a missing value in a linear model by regressing on the non-missing value most highly correlated with the missing variable.

There is a considerable amount of literature on missing value estimation. In the following only a brief look at some of the main issues is included. In the comparative work carried out by Titterington et al (1981) the data were assumed to be *missing at random* within each prognostic category. Here *at random* is according to the following definition given by Little (1978) as being equivalent to that given by Rubin (1976).

Definition 3-1: *Let the q-variate data matrix* $\mathbf{X}$ *with n observations be given by* $\mathbf{X}=\{x_{ij}\}$ *and let* $\mathbf{R}$ *be the random matrix* $\mathbf{R}=\{r_{ij}\}$ *with* $r_{ij}=0$ *or* 1 *according to whether* x_{ij} *is missing or observed. Then any missing values are missing at random if the distribution function of the conditional distribution of* $\mathbf{R}$ *given* $\mathbf{X}$ *is functionally independent of the missing values. In particular, the probability that a value* x_{ij} *is observed must not depend on the values of an observed variable* x_{ik}.

Rubin (1976) points out that this definition constitutes the weakest definition of *missing at random* which allows one to ignore the mechanism generating the missing values which can be fairly unrealistic in some applications. Arminger and Sobel (1990) and Rehm (1990) take issue with conventional methods for estimating missing data such as given in the first paragraph of this subsection. They claim that these may be shown to be either statistically inefficient or have unknown statistical properties. The statistical basis for this argument is laid out in Arminger and Sobel (1990). If missing values are estimated from the

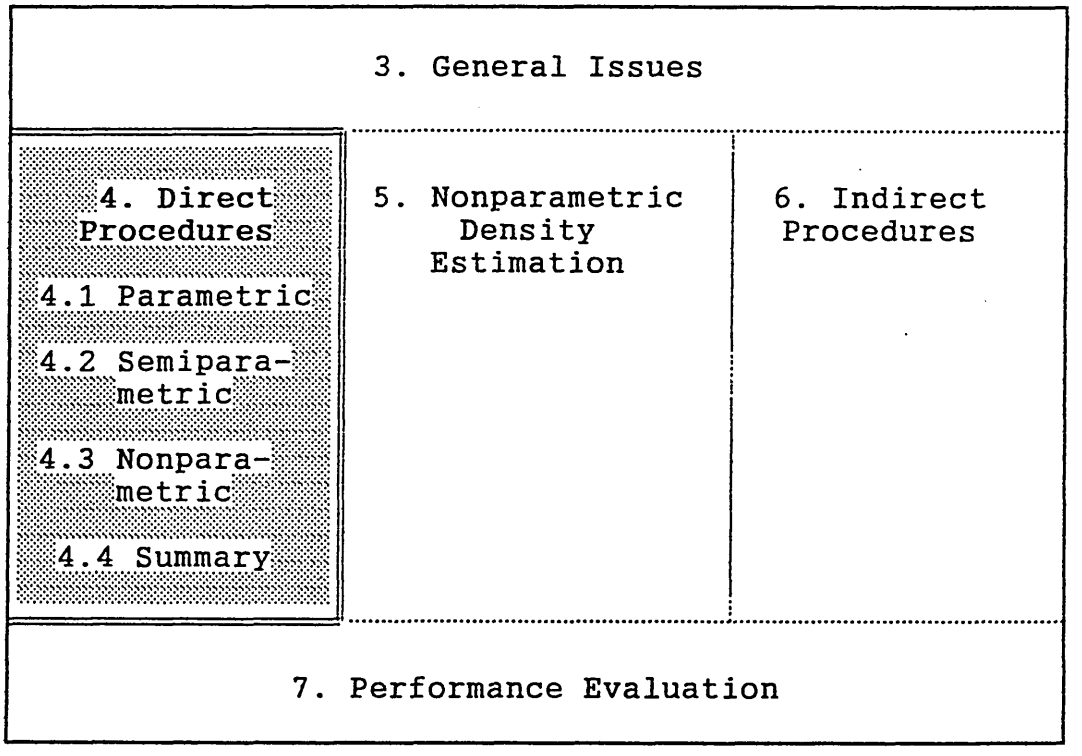
data then the expected overall variance, as measured by the residual sum of squares in regression analysis for example, will be underestimated due to the sample dependence of the missing value estimate. In the view of the authors no well established procedures for correcting for this bias in variance estimates exist to date. In the absence of such general guides it will seem to be most pragmatic to consider only datasets with roughly similar proportions of missing data in comparative work. As the absolute size of the residual variance is of lesser importance in comparative work this bias may be taken as a fair trade off.

3.4 Summary

Discriminant analysis, especially for discrete data situations, draws heavily on five major areas of statistics: *model assumptions, density estimation, construction of distance measures, performance assessment* and the *analysis of posterior probabilities.* The review (part II) is largely structured corresponding to these different fields. There is a strong emphasis in the literature on the behaviour of error rates and on the concept of optimality of a discriminant rule. By contrast there is comparatively little work on selection guides for discrete discriminants. There is also not much work reporting use of posterior distributions in constructing performance criteria. Special issues not considered central to the research as stated at the end of the introduction (part I) are *selection of variables, transformation of variables* and *missing data estimation.* The work that is most relevant to the study of general issues in discriminant analysis includes the 8 papers by Fix & Hodges (1951), Hills (1966), Lachenbruch (1975), Victor (1976), Titterington et al (1981), Hora & Wilcox (1982), Goldstein and Efron (1983) & Dillon (1984).

I: INTRODUCTION

II: REVIEW



III: METHOD

IV RESULTS

V: DISCUSSION

Because of their different modes of operation it proves useful to distinguish two general classes of procedures for discriminant analysis subsequently referred to as *direct* and *indirect* procedures. The latter are discussed in chapter 6. Direct procedures are based on statistical data models that provide direct expressions for the posterior probabilities of population membership. Generally population specific densities will also be estimable.

When prior probabilities are assumed to have prior distribution $f(\pi_i) = P_r(\Pi_i=\pi_i)$ where Π_i is a random variable the posterior density function $f(\pi_i|\mathbf{x})$ is proportional to the likelihood times the prior density

$$f(\pi_i|\mathbf{x}) \propto f(\pi_i)L(\pi_i;\mathbf{x}) \tag{4-1}$$

or

$$f(\pi_i|\mathbf{x}) \propto f(\pi_i)f(\mathbf{x}|\pi_i). \tag{4-2}$$

In the present thesis all examples of datasets analysed in part IV of the text are discussed under the assumption that the prior vector $\pi = (\pi_1, \ldots, \pi_g)'$ defining probability of population membership is specified, i.e. the π_i are not treated as observations of random variables Π_i. The allocation rules will thus depend on the likelihood ratio weighted by corresponding fixed priors. For two groups this is

$$LR = \frac{\pi_1 f(\mathbf{x}|\Pi_1)}{\pi_2 f(\mathbf{x}|\Pi_2)} . \tag{4-3}$$

This approach is equivalent to discussing Bayesian discriminant analysis under the condition of a uniform prior distribution. Therefore, to avoid confusion with a fully Bayesian approach to discriminant analysis, the special case of Bayesian discriminant procedures with a uniform prior is thus subsequently referred to as the class

of *direct* procedures. Note that Bayesian approaches are not necessarily exclusive to direct procedures. A typical example of an indirect (the distributional distance) procedure with a Bayesian approach is given in Lack (1987) and is also discussed in chapter 6. The distinction between Bayesian and non-Bayesian discriminant procedures is not treated unanimously. Mardia, Kent and Bibby (1979, chapter 11) for instance consider direct procedures in the above sense as standard Bayesian. By contrast Dunsmore (1966) and Aitchison and Dunsmore (1975, chapter 1) emphasise the importance of the prior distribution.

As the methodology to be developed will to a large extent focus on direct procedures greater detail is given for this class of procedures. Direct procedures themselves may be divided again into *parametric* (section 4.1), *semiparametric* (section 4.2) and *nonparametric* procedures (section 4.3). The topic of *nonparametric density estimation* is so central to *nonparametric discriminant procedures* that it is treated separately in chapter 5.

Figure 4-1 gives an overview of major papers on comparative work and discriminant procedures (direct and indirect) suitable for discrete data including references covered both in this chapter and in chapter 6.

distance methods	Matusita	1955
	Matusita	1956
	Goldstein & Dillon	1978
	Moore	1982
	Pridmore	1985
linear functions	Fisher	1936
	Smith	1947
	Raveh	1989
logistic	Cox	1972
	Anderson	1975
	Feldmann et al	1981
	Bull & Donner	1987
	Feldmann	1987
	Campbell & Donner	1989
interaction	Bahadur	1961
	Lancaster	1969
	Cox	1972
	Moore	1973
	Zentgraf	1975
	Goldstein & Dillon	1978
recursive partitioning	Sonquist & Morgan	1964
	Breiman et al	1984
	Loh & Vanichsetakul	1988
neural networks	Hertz et al	1991
	Ripley	1994
comparative studies	Gilbert	1968
	Moore	1973
	Goldstein & Dillon	1978
	Titterington et al	1981
	Loh & Vanichsetakul	1988

Figure 4-1: References on discrete discriminants

Figure 4-1 shows major references on discrete discriminant analysis including year of publication. Other typically more recent references not shown relate to work on mixed discrete and continuous distributions. A typical example of this is the *location model* reported by Krzanowski (1982) and Vlachonikolis (1990). References to indirect procedures such as recursive partitioning, pattern recognition and neural networks are typically more recent.

Discriminant procedures are often divided into *parametric* and *nonparametric* procedures depending on whether a data model exists or not. In the former case - for instance if the data stem from a multivariate trinomial distribution -

the given distribution function will be characterised by parameters that are either given or require estimation. In the latter case only the data are given with no indication as to how they were generated. Nonparametric procedures provide local density estimates. An example of such a density estimation based procedure is the nearest neighbour discriminant (Fix and Hodges, 1951; Devroye and Wagner, 1982; Silverman and Jones, 1989).

Logistic discriminant procedures (Anderson, 1982; Albert and Lesaffre, 1986) do not fit easily into either class and have therefore been classed separately as *semiparametric* or *partially parametric* procedures by some authors (eg, McLachlan, 1992). This distinction has been adopted in the present text. More modern discriminant procedures such as *classification trees* (Breimann, Olshon, Friedman and Stone, 1984; Loh and Vanichsetakul, 1988) or *neural networks* (Ripley, 1993, 1994) use iterative steps to solve the discriminant problem and these are referred to as procedures based on *recursive partitioning*. All procedures including indirect procedures considered in this review chapter are summarised graphically in relation to their class in figure 4-2.

Distance procedures have been extensively described by Goldstein and Dillon (1978) who generalised Matusita's (1955) population distance. Later work includes comparative studies of the *generalised distance* and other distance measures (Moore, 1982; Pridmore, 1985; Lack, 1987).

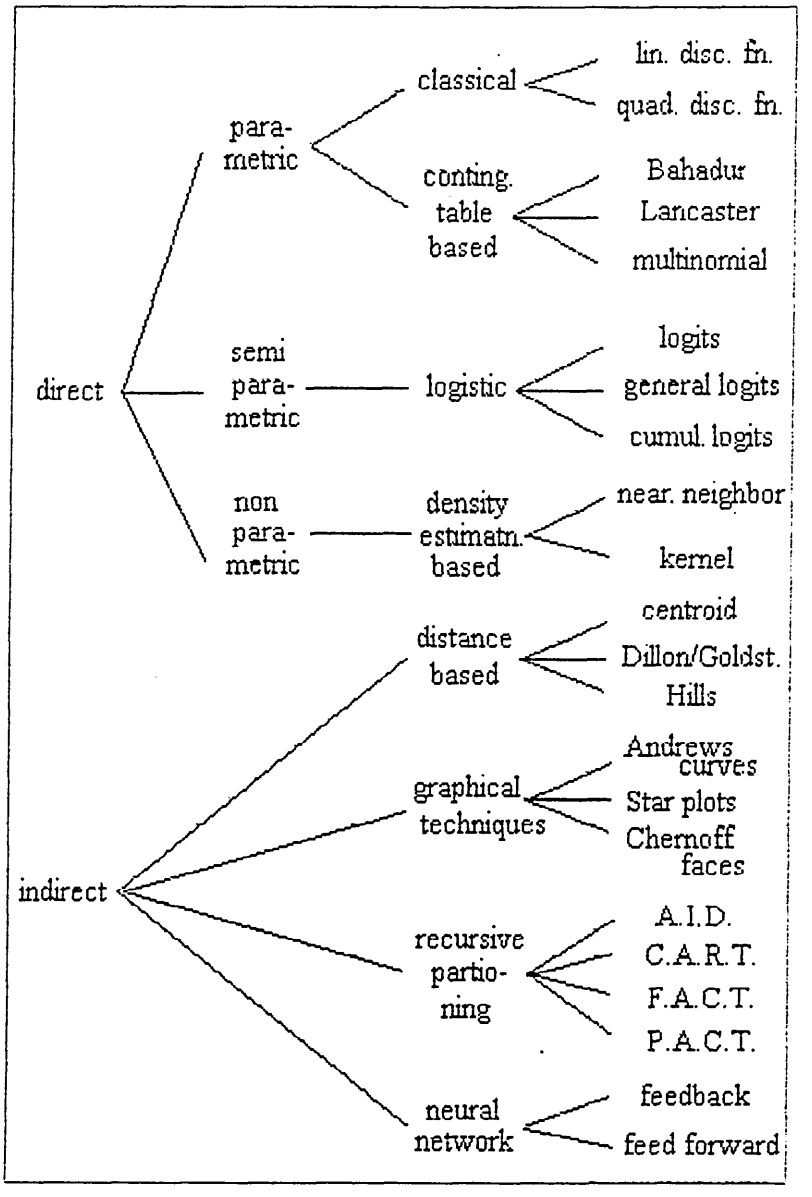


Figure 4-2: Overview of discriminant procedures

Figure 4-2 shows a schematic overview of procedures for discrete discriminant analysis. The procedures above apply when the samples contain labelled observations i.e., when population membership is known. In the field of pattern recognition this condition is commonly termed *supervised learning*. The procedures are divided depending on whether posterior probabilities are used for allocation (*direct procedures*) or whether allocations are made without posterior probabilities (*indirect procedures*).

The main *parametric procedures* most widely used are the *linear discriminant function* due to Fisher (1936) and its extension to the *quadratic discriminant function* for heteroscedastic covariance matrices. Equally important, though far less frequently applied, is the class of procedures based on *interaction models* ranging from the *independent model*, through models such as based on the *Bahadur* representation of multivariate dichotomous data suggested by Bahadur (1961a, 1961b), or the equivalent based on *Lancaster models* for polychotomous data due to Lancaster (1969) to the *multinomial procedure*. Common to all parametric procedures is that probability models for the conditional and joint densities are the starting point from which suitable estimates of the prior probabilities of population membership and posterior probabilities are then derived.

Performance of the *linear discriminant* (Fisher, 1936) with discrete data has been extensively investigated. Major early comparative studies are reported by Cochran and Hopkins (1961), Gilbert (1968), Moore II (1973), and Goldstein and Dillon (1978). Though theoretically inappropriate in the case of discrete data the linear discriminant generally performs remarkably well (Goldstein & Dillon, 1978; Hand, 1983). Exceptions to this robust behaviour generally exist when the data are extremely non-normal or when there is a substantial degree of correlation between the variables (Gilbert, 1968; Moore, 1973).

By contrast when applied to continuous data the linear discriminant has the advantage that it is easy to apply, is implemented in standard statistical software packages, has fast execution speed due to comparatively few parameters requiring estimation and above all exhibits a high degree of robustness. The fact that this last feature is no longer always true in the case of discrete data highlights again

the need for selection guides for discriminant procedures for discrete data.

One explanation for the poor performance of the linear discriminant in the presence of correlations is that the ratio of the likelihoods undergoes a *reversal* (Moore II, 1973), i.e., is nonmonotonic. In the case of two binary variables a reversal will occur whenever the states at the two diagonal points *(0,0)* and *(1,1)* have greater frequency in one population while the states at *(0,1)* and *(1,0)* are more frequent in the other. Here a linear discriminant does not perform satisfactorily (table 4.1-1, figure 4.1-1).

$state_j$	n_{1j}	n_{2j}	$\ln((n_{1j}/n_{2j})$
(0.0)	*10*	*90*	*-2.20*
(0.1)	*90*	*10*	*2.20*
(1.0)	*90*	*10*	*2.20*
(1.1)	*10*	*90*	*-2.20*
n_{i+}	*200*	*200*	

Table 4.1-1: Hypothetical *4*-state example data

Table 4.1-1 shows a hypothetical *4*-state example for two bivariate dichotomous populations located at opposite corners of a square in the cartesian plane. The final column shows the typical *reversal of the loglikelihood ratio.* In this case complete separation is not possible with the linear discriminant nor with a simple curvilinear discriminant function.

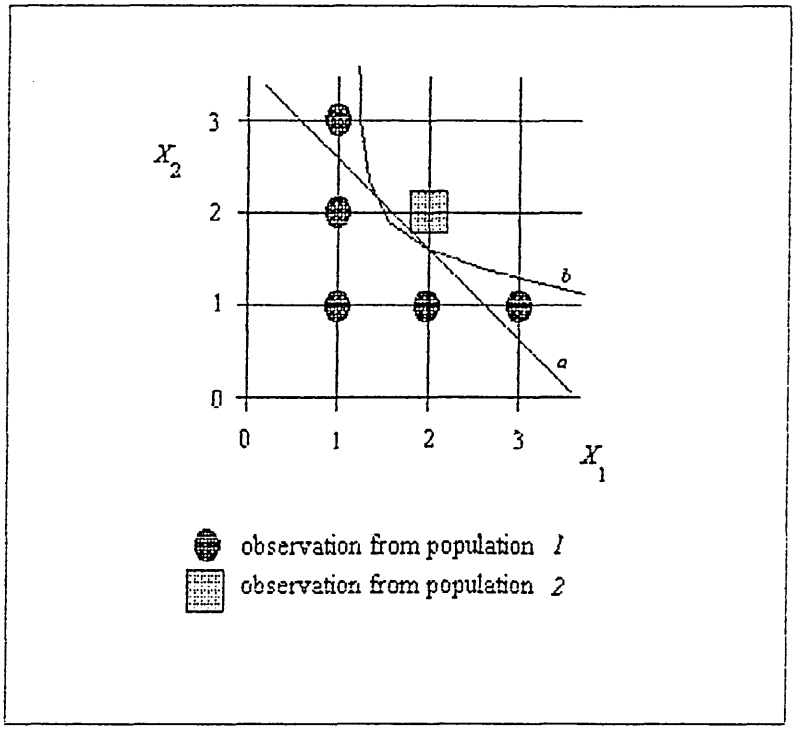


Figure 4.1-1: Curvilinear separation lines

Figure 4.1-1 shows a hypothetical *6*-state example of two populations distributed in the cartesian plane with X_1 and X_2 discrete. Population *2* is exclusively located at $(X_1=2; X_2=2)$. Linear separation lines (*a*) are inferior. Complete separation is possible only with curvilinear separation lines (*b*) such as used in *quadratic discriminant analysis*.

In other approaches the interactive data structure is modelled directly. A method for multivariate binary data is given by Bahadur (1961a, 1961b) and for the more general discrete case by Lancaster (1969). Both express the conditional distribution solely in terms of means and correlations. Moore II (1973), Zentgraf (1975), Goldstein and Dillon (1978) report major comparative studies of *interaction models*. The following details the main points of this model family.

Assume that for the i^{th} of g populations q-variate data is available with $\mathbf{X}_i=(X_{i1},X_{i2},\ldots,X_{iq})'$, $i=1,\ldots,g$. The simplest discrete data structure is dichotomous such that

for population Π_i each component X_{ik} $(k=1,\ldots,q)$ is a *Bernoulli* random variable with $\Pr\{X_{ik}=1\}=p_{ik}$. Thus the vector $\mathbf{X}_i$ is an element in q-variate Bernoulli space: $\mathbf{X}_i \in \mathbb{B}^q$. Data of this type frequently arise in social science surveys or medical research. Positive and negative responses to questions or presence and absence of symptoms are recorded. When $q=2$ then 4 distinct states result for $\mathbf{X}_i$ namely $(0,0)', (0,1)', (1,0)'$ and $(1,1)'$. For q-variate dichotomous data the number of distinct states is 2^q. The simplest statistical model of multivariate data structure is given when independence between the components X_{ik} is assumed such that $COV(X_{ik}, X_{il})=0$. This makes estimation of the joint density function easy as it is simply the product of the marginal densities

$$f_i(\mathbf{x}_j) = \prod_{k=1}^{q} f_{ik}(x_{jk}) \qquad (4.1\text{-}1)$$

where x_{jk} is the j^{th} observation of the k^{th} component of the q-variate random variable $\mathbf{X}$ and f_{ik} is the corresponding marginal density function in the i^{th} population. Bernoulli variables take on the values 0 or 1. So if X_k is a random component of $\mathbf{X}$ from Π_i then $p_{ik} = \Pr\{X_k=1 | \mathbf{X}\in\Pi_i\}$. Thus the likelihood for Π_i given a sample of n_i observations is

$$l_i(\mathbf{x}) = \prod_{j=1}^{n_i} f_i(\mathbf{x}_j) = \prod_{j=1}^{n_i} \prod_{k=1}^{q} f_{ik}(x_{jk}) \qquad (4.1\text{-}2)$$

or

$$l_i(\mathbf{x}) = \prod_{j=1}^{n_i} \prod_{k=1}^{q} p_{ik}^{\,x_{jk}} (1 - p_{ik})^{(1 - x_{jk})} . \qquad (4.1\text{-}3)$$

For $g=2$ and equal priors the *maximum likelihood* discriminant rule under the independence model is

therefore: "*allocate a new observation* x_0[6] *with unknown population membership to* Π_1 *if* $l_1(x_0)/l_2(x_0)$ *exceeds unity.*" To apply this rule parameters p_{ik} have to be estimated as they are generally unknown. The maximum likelihood estimates turn out to be simply the observed relative marginal frequencies.

The independence model is a restricted model. Generally it is realistic to assume some degree of correlation between the predictor variables. Bahadur (1961a) suggested a model for representing multivariate binary data incorporating such interactive structure as a probability density using the transformation

$$Z_{ik} = \frac{X_{ik} - p_{ik}}{\sqrt{p_{ik}(1-p_{ik})}} \qquad (4.1\text{-}4)$$

with p_{ik} defined as above. Let correlational terms be defined by the conditional expectations

$$\rho_{i,k\ldots m} = E\left[X_{ik}\, X_{il} \ldots X_{im} \mid \mathbf{X}\in\Pi_i\right] \qquad (4.1\text{-}5)$$

with the second index running from k through m. For a saturated model the second index runs from k through to q where q is the total number of predictor variables. Lazarsfeld (1956, 1961) and Bahadur (1961a) both independently showed that $f_i(\mathbf{x}_j)$ could be reparametrised as

$$f_i(\mathbf{x}_j) = \prod_{k=1}^{q} p_{ik}^{x_{jk}} (1 - p_{ik})^{(1 - x_{jk})} \qquad (4.1\text{-}6)$$

$$\left[1 + \sum_{k<l} \rho_{i,kl}\, Z_{ik}\, Z_{il} + \ldots + \sum_{k<\ldots<m} \rho_{i,k\ldots m}\, Z_{ik}\, Z_{im} \right] .$$

Thus (4.1-6) provides an expression for multivariate binary data solely in terms of means p_{ik} and correlations $\rho_{i,kl},\ldots,\rho_{i,k\ldots m}$. The set of parameters $\{p_{ik}\}$ may be seen

[6] where $\mathbf{X}_0$ has not been included in estimation of the rule

as means because the original variables X_{ijk} are Bernoulli variables. By including only correlational terms between two variables such as $\rho_{i,kl}$, second order models may be defined. A first order model has no correlational terms and may immediately be seen to be equivalent to the independence model which is the same as (4.1-6) without the sum in the square brackets. When the maximum correlational terms are included a saturated model results where each of the cells in the multidimensional contingency table has its own parameter. This corresponds to the *full multinomial* model which is generally written more simply as

$$\begin{aligned} \Pr\{X_i=x\} &= \Pr\{X_{i1}=x_1,\ldots,X_{iq}=x_q \mid X\in\Pi_i\} \\ &= p_{i,k\ldots m} \cdot \end{aligned} \qquad (4.1\text{-}7)$$

The number of parameters to be estimated in the multinomial model can be substantial as it is given by $\prod_{k=1}^{q} l_k$ where l_k is the number of levels of different values that the component X_k can take on.

Independence, Bahadur and full multinomial models may be seen to belong to the family of *interaction models* characterised by varying degrees of correlational order among the variables. Zentgraf (1975) and Trampisch (1976) suggested an extension of the Bahadur model using Lancaster's (1969) definition of higher order interaction models for discrete data. The Lancaster model is

$$\begin{aligned} p_{j_1\ldots j_q} = \sum_{v=2}^{s} (-1)^{s-v} \binom{q-v-1}{s-v} \\ \cdot \sum_{C_q^v} p_{j_{k_1}\ldots j_{k_v}}^{(k_1\ldots k_v)} \, p_{j_{m_1}}^{(m_1)} \ldots p_{j_{m_{q-v}}}^{(m_{q-v})} \\ + (-1)^s \left[\binom{q-1}{s} - q\binom{q-2}{s-1}\right] \prod_{m=1}^{q} p_{j_m}^{(m)} \end{aligned} \qquad (4.1\text{-}8)$$

where C_q^v stands for the set of all combinations of v elements out of $\{1,\ldots,q\}$. Here $p_{j_1\ldots j_q}$ is a cell probability, i.e., the joint probability for the variables X_m to have outcomes j_m respectively with $m = 1,\ldots,q$. Further, $p_j^{(m)}$ stands for the probability of the j^{th} outcome of X_m, $p_{jk}^{(mn)}$ for the probability of the joint occurrence of the j^{th} outcome of X_m and the k^{th} occurrence of X_n, etc. This formulation corresponds to the *index-dot* notation (Zentgraf, 1975) with the modification that only those indices which are not substituted by a dot are given and their position is marked by a superscript. Zentgraf (1975) gives further expressions of (4.1-8) when higher order interactions vanish, in particular for $s = 2$.

The main advantage of (4.1-8) is based on the fact that it is possible to calculate the cell probabilities explicitly by use of marginal probabilities of first up to s^{th} order if disappearance of interactions higher than s^{th} order is assumed. The relative frequencies may be used as simple estimates of the marginal probabilities. These estimates are stable for sufficiently small s and hence also for the estimates of the cell probabilities. For dichotomous variables (4.1-8) simplifies to (4.1-6) above by either using the symmetric parameters introduced by Lazarsfeld (1961) or equivalently using the n^{th} order correlation parameters introduced by Bahadur (1961a). The same result will be achieved with these representations since the vanishing of Lancaster's s^{th} order interactions, the disappearance of Lazarsfeld's $(s+1)^{th}$ order symmetric parameters and the vanishing of Bahadur's $(s+1)^{th}$ order correlations are all equivalent. Goldstein and Dillon (1978) also showed that Bahadur models for multivariate binary data are contained in Lancaster models for discrete data.

Lancaster models complete the family of interactive data models for discriminant analysis. Interactive discriminant procedures of the above type have been extensively studied by Bahadur (1961a, 1961b), Lazarsfeld (1961), Gilbert

(1968), Moore (1973), Victor et al (1974), Trampisch (1976) Goldstein and Dillon (1978), and more recently also by Moore (1982), Pridmore (1985) and McLachlan (1992). The skill in application lies in selecting the optimal interactive order. While the independence assumption may not do justice to, and therefore inadequately represent, data containing interactions, the full multinomial model may occasionally be over specified.

In practical applications a balance will have to be struck between choice of a parsimonious model, of which the independence model is an extreme case and on the other hand the theoretically appropriate model of which the fully saturated multinomial model is also an extreme case. Whilst the independence model generally leads to few but stable parameter estimates the full multinomial model will exhibit a large number of parameters that may tend to be unstable. The range of data models from independence to full multinomial may be seen to constitute a *family* of procedures. For a general discussion of these considerations see for instance Bishop, Fienberg and Holland (1975) or Victor (1976) for the problem of choice of discriminant procedure from a family of related procedures.

4.2 Semiparametric procedures

The class of *semiparametric procedures* embraces all variants subsumed under the heading of *logistic discriminant analysis*. The essential difference to parametric procedures is that *suitable ratios* of the posterior probabilities of population membership can now be modelled *directly*. Knowledge of the group conditional densities is *not* required. Logistic discriminant analysis has three chief advantages. Firstly, the distributional assumptions are relaxed such that the procedures may be applied to continuous as well as discrete data. Secondly, linear discriminant analysis can be shown to be a special

case of the more general logistic approach. Thirdly, extensions of the basic logistic model allow the use of information contained in ordinal explanatory (predictor) variables as well as the modelling of ordinal response categories for the dependent variable.

The logistic model (Cox, 1972) assumes that the logit of the binary response can be expressed as a linear combination of variables. These may be continuous, nominal or even dichotomous. The logistic can also copy with *reversals* in the likelihood ratios. Extensions of the basic logistic model involve expressing the logit of the binary response as a mixture of linear terms as well as quadratic terms (Anderson, 1975), ordered outcome categories for the dependent variable (Feldmann et al, 1981; Feldmann, 1987) and discrimination between three or more groups (Bull and Donner, 1987).

Generally, the results reported in the literature suggest that logistic discrimination is preferable to other widely used methods for multiple group classification with non-normal data, and is comparable to classification by multiple linear discrimination with normal data (Baron 1991). Procedures based on the *logistic model* are probably the most widely used alternative to the linear discriminant. This is largely due to the ready availability of easily applied logistic discriminant procedures in contemporary statistical software packages such as *SAS*, *SPSS*, *GENSTAT* or *BMDP*. By contrast the nonparametric nearest neighbour or kernel density based procedures discussed in chapter 5 are much less popular because they require in addition the estimation of further smoothing parameters. This makes such procedures less accessible for ready use especially by non-statisticians.

Loglinear models due to Birch (1963) and developed by Nelder and Wedderburn (1972) are well established models for discrete data. Comprehensive applications are given in Bishop, Fienberg and Holland (1975) and Fienberg (1980) for example. The model expresses the logarithm of all state probabilities including the factor containing group membership, x_1, as a linear combination of main effects, $(-1)^{x_j}\alpha_j$, and interactions, $(-1)^{x_j+x_k}\alpha_{jk}$, where $X \in \{0,1\}$.

Suppose that

$$\log f(\mathbf{x}) = \alpha + \sum_{j=1}^{q} (-1)^{x_j}\,\alpha_j + \sum_{j<k} (-1)^{x_j+x_k}\,\alpha_{jk} + \ldots$$

$$\ldots + (-1)^{x_1+x_2+\ldots+x_q}\,\alpha_{12\ldots q} \qquad (4.2\text{-}1)$$

where α is an overall effect, α_j is the main effect due to X_j and α_{jk} is the respective interaction effect between X_j and X_k and so on. The discrete density is specified by estimating respective main effect and interactive terms in (4.2-1). The vector $\mathbf{x}$ includes the factor containing group membership, x_1, say.

In application to discriminant analysis significant interaction effects including X_1 will indicate good discrimination. Using the notation of Bishop, Fienberg and Holland (1975) and Fienberg (1980), the saturated model (4.2-1) may be written in the $q=3$ variable situation with X_1 containing group membership, $i=1,\ldots,g$, as

$$\ln p_{ijk} = u+u_1(i)+u_2(j)+u_3(k)+u_{12}(ij)+u_{13}(ik) +u_{23}(jk)+u_{123}(ijk) \qquad (4.2\text{-}2)$$

where

$$\Sigma_i u_1(i)=\Sigma_j u_2(j)=\Sigma_k u_3(k)=0,$$

$$\Sigma_i u_{12}(ij)=\Sigma_j u_{12}(ij)=\Sigma_i u_{13}(ik)=\Sigma_k u_{13}(ik)=\Sigma_j u_{23}(jk)=\Sigma_k u_{23}(jk)=0 \text{ and}$$

$$\Sigma_i u_{123}(ijk)=\Sigma_j u_{123}(ijk)=\Sigma_k u_{123}(ijk)=0.$$

Expression (4.2-2) corresponds in the usual notation to the discrimination problem of allocation of bivariate discrete objects to one of g populations.

4.2.2 Logistic models

Loglinear models and logistic models may be shown to be equivalent in the case of a dichotomous response variable. If, as applies in this case, the number of groups, g, is set at *2* in (4.2-2) and (bivariate) state probabilities of group membership, $p_{jk} = \Pr\{X_1=1 \mid X_2=j \cap X_3=k\}$ where $X_2, X_3 \in \{-1,+1\}$, are expressed as dependent variables the logistic model obtains

$$\text{logit}(p_{jk}) = \ln\left(\frac{p_{jk}}{1-p_{jk}}\right) = \alpha' x$$

$$= \alpha_0+\alpha_1 x_1+\alpha_2 x_2+\alpha_{1\ 2} x_1 x_2 \qquad (\ 4.2\text{-}3)$$

with α_{12} corresponding to the interaction effect between X_1 and X_2. The logistic model may also be derived from normally distributed observations in populations Π_1 and Π_2, where the posterior probabilities following Bayes' theorem may be written as

$$\Pr(\Pi_1|\mathbf{x}) = e^{\alpha' \mathbf{x}} \times \Pr(\Pi_2|\mathbf{x}) = \frac{e^{\alpha' \mathbf{x}}}{1 + e^{\alpha' \mathbf{x}}} \qquad (\ 4.2\text{-}4)$$

$$\text{where} \qquad \Pr(\Pi_2|\mathbf{x}) = \frac{1}{1 + e^{\alpha' \mathbf{x}}} .$$

This may be seen by writing the posterior probability $Pr(\Pi_1|\mathbf{x})$ as

$$\Pr(\Pi_1|\mathbf{x}) = \frac{\Pr(\mathbf{x}|\Pi_1)\Pr(\Pi_1)}{\Pr(\mathbf{x}|\Pi_1)\Pr(\Pi_1) + \Pr(\mathbf{x}|\Pi_2)\Pr(\Pi_2)} \qquad (\ 4.2\text{-}5)$$

which is equivalent to

$$Pr(\Pi_1|x) = \frac{1}{1 + \frac{Pr(x|\Pi_2)\ \pi_2}{Pr(x|\Pi_1)\ \pi_1}} \ . \qquad (4.2\text{-}6)$$

The fraction in the denominator of 4.2-6 is the ratio of multivariate normal group conditional densities and in the case of homoscedasticity its logarithm may be written in matrix notation as

$$(x-\mu_2)'\Sigma^{-1}(x-\mu_2) - (x-\mu_1)'\Sigma^{-1}(x-\mu_1) \qquad (4.2\text{-}7)$$

which is the well known linear discriminant function

$$-\frac{1}{2}\ (\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) + (\mu_1-\mu_2)'\Sigma^{-1}x \ . \qquad (4.2\text{-}8)$$

Upon resubstitution of $\beta_0 = -\frac{1}{2}\ (\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$ and $\beta = (\mu_1-\mu_2)'\Sigma^{-1}$ and suitable reparametrisation equations 4.2-4 above are obtained.

The parameter vector α has dimension $(q+1)$ and depends on the mean vector and covariance matrix of X. If this dependence assumption is relaxed and α is regarded as an independent parameter, then expression (4.2-4) denotes the logistic discrimination classification procedure which may also be applied to a wide class of non-normal distributions. As logistic discrimination contains the linear discriminant function as a special case it may also be applied in all situations in which the latter leads to good results. An object with unknown class membership Π_i is usually allocated to that class giving the larger posterior probability, $Pr(\Pi_i|x)$.

An extension of the logistic discrimination procedure to g populations has been given by Anderson (1982).

$$\Pr(\Pi_i|\mathbf{x}) = e^{\alpha_i' \mathbf{x}} \times \Pr(\Pi_g|\mathbf{x})$$

$$\Pr(\Pi_g|\mathbf{x}) = \frac{1}{1 + \sum_{s=1}^{g-1} e^{\alpha_s' \mathbf{x}}} \qquad (4.2\text{-}9)$$

where $\alpha_i' = (\alpha_{io}, \alpha_{i1}, \ldots, \alpha_i)$ $(i=1,\ldots,g-1)$, and for ordered categories by Feldmann et al (1981) and Feldmann (1987). Further extensions of the linear exponents $\alpha'\mathbf{x}$ in (4.2-4) and (4.2-5) are given in Anderson's (1975) quadratic logistic discrimination where the exponent in (4.2-9) takes the form $\alpha_i'\mathbf{x} + \beta'\mathbf{x}\mathbf{x}'\beta$, the second term being the quadratic component.

The extension of the logistic procedure to more than *2* populations is achieved by using *generalised logits*. Let the probabilities for the j^{th} state be denoted by p_{1j}, p_{2j}, ..., p_{gj}. Then generalised logits are obtained by taking the logarithms of the ratios of two probabilities where the denominator of each ratio corresponds to the last observed level of the dependent variable. Thus *g-1* ratios are obtained:

$$\begin{aligned} \text{logit}\ (p_{1j}) &= \ln\ (p_{1j}/p_{gj}) \\ \text{logit}\ (p_{2j}) &= \ln\ (p_{2j}/p_{gj}) \\ &\vdots \\ \text{logit}(p_{g-1j}) &= \ln\ (p_{g-\ 1j}/p_{gj}) \end{aligned} \qquad (4.2\text{-}10)$$

When the response variable is ordinal the logistic model offers a further refinement in terms of *cumulative logits*. Ordinal response is not unusual. In clinical trials research a discriminant approach may be required in order to shed light on differences between patients whose post treatment outcome was classed as unchanged, moderate or satisfactory. Again logarithms of the ratios of two probabilities for a given state *j* are taken, yet the denominator of the k^{th} ratio is the cumulative probability c_{kj} corresponding to the k^{th} level of the independent variable:

$$\text{i.e. if } Y \in \{1,\ldots,g\} \text{ then } c_{kj} = \sum_{i=1}^{k} p_{ij}.$$

The numerator in each case is $1-c_{kj}$. Thus ratios are obtained for cumulative logits:

$$\begin{aligned}
&\text{cum logit } (p_{1j}) = \ln((1-p_{1j})/p_{1j}) \\
&\text{cum logit } (p_{2j}) = \ln((1-p_{1j}-p_{2j})/(p_{1j}+p_{2j})) \\
&\qquad\qquad \vdots \qquad\qquad\qquad\qquad\qquad\qquad (4.2\text{-}11) \\
&\text{cum logit } (p_{g-1j}) = \ln(p_{gj}/(p_{1j}+p_{2j}+\ldots+p_{g-1j})).
\end{aligned}$$

Advantages in the logistic model clearly lie in its applicability to non-normal data and in the comparatively few parameters that require estimation as well as the unrestrictive modelling of interactions by specifying individual effects. The success of the logistic approach to discrimination will depend, however, on the extent to which the assumption that the logit may be modelled as a linear combination of effects as in (4.2-3). An application of (4.2-3) is given in Lack (1987).

4.3 Nonparametric procedures

At the heart of the *nonparametric procedures* lies the problem of density estimation (further information on nonparametric density estimation is given in chapter 5) in the absence of prior distributional information. Once estimates have been obtained the discriminant rule is constructed using the standard likelihood ratio approach. The unknown densities may be estimated by obtaining suitable estimates for multinomial cell probabilities. These are either used directly or may be smoothed *orthogonal series* (Ott and Kronmal, 1976). Alternatively *nearest neighbour* (Fix and Hodges, 1951; Hand, 1982) or *kernel density estimation* (Rosenblatt, 1956; Parzen, 1962;

Titterington, 1980) techniques may be applied. Kernel density estimates are classed as either *fixed* or *adaptive* in reference to the way in which the *smoothing parameter*, commonly called λ, is determined. Kernel functions may differ slightly depending on whether they are used to estimate multivariate binary data, categorical data or ordinal data.

4.3.1 Nearest neighbour procedures

Nearest neighbour procedures were first applied to discriminant analysis by Fix and Hodges (1951) and have subsequently been employed by Loftsgaarden and Quesenberry (1965), Hills (1967), Trampisch (1976), Hall (1981b), and Hand (1982) among others. There have been suggestions as to how the smoothing parameter, k (see below) should best be estimated. No clear rule emerges, however, with the conclusion that the choice of k is relatively unimportant. The advantages of nearest neighbour techniques are to be seen in the fact that no prior assumptions are required and that application is relatively easy.

Let $d_{i,k}(\mathbf{x})$ be the euclidean distance from the k^{th} nearest point among $\mathbf{x}_{ij}$, $(j=1,\ldots,n_i;\ i=1,\ldots,g)$. Then the nearest neighbour density estimate of order k in the i^{th} population is defined by

$$\hat{f}_i^{(NN)}(\mathbf{x}) = \frac{k}{n_i v_q \{d_{i,k}(\mathbf{x})\}^q} \qquad (4.3\text{-}1)$$

where v_q is the volume of the unit sphere in q dimensions (so that $v_1=2$, $v_2=\pi$, $v_3=(4/3)\pi$, etc.). Choice of the parameter k is crucial as small values lead to locally sensitive density estimates while large values of k will tend to smooth the estimates $\hat{f}_i^{(NN)}(\mathbf{x})$. For $g=2$ populations the nearest neighbour discriminant rule is based on the likelihood ratio

$$LR = \frac{\hat{f}_1^{(NN)}(x)}{\hat{f}_2^{(NN)}(x)} \qquad (4.3\text{-}2)$$

A new observation x is allocated to population Π_1 when the likelihood ratio *LR* exceeds unity. In the rare event of *LR=1* allocation of x is at random.

4.3.2 Methods based on kernel density estimation

Kernel estimates of the unknown density functions have been employed, among others, by Hills (1967), in application to multivariate binary data by Aitchison and Aitken (1976), in a more general form applied to mixed data by Titterington et al (1981) and with improved estimation of λ in application to binary data by Hall (1981b).

Aitchison and Aitken apply expression (5.1-3) from chapter 5 to a multivariate (*q=10*) binary dataset with $n_1=40$ training and $n_2=41$ test cases reported by Anderson et al (1972) of patients suffering from *keratoconjunctivitis sicca*, (*KCS*), and suitable non-*KCS* controls. They compare the kernel method with the independence model, the linear discriminant function, the loglinear and logistic models and a nearest neighbour technique. The authors found that discrimination by the kernel method is perfect for the test set of cases. Hall, (1981) employs the same kernel estimator but estimates the smoothing parameter, λ, using a crossvalidated maximum likelihood approach and reanalyses the *KCS* data. The smoothing parameter, λ, is estimated such that a global function of the mean squared error is minimised. This has advantages when the number of cells in the multinomial sample is large.

In summary, the results on using kernel density estimators in discrete discriminant analysis are scanty and sometimes even equivocal, largely due to few applications and also because of a marked lack of comprehensive sampling

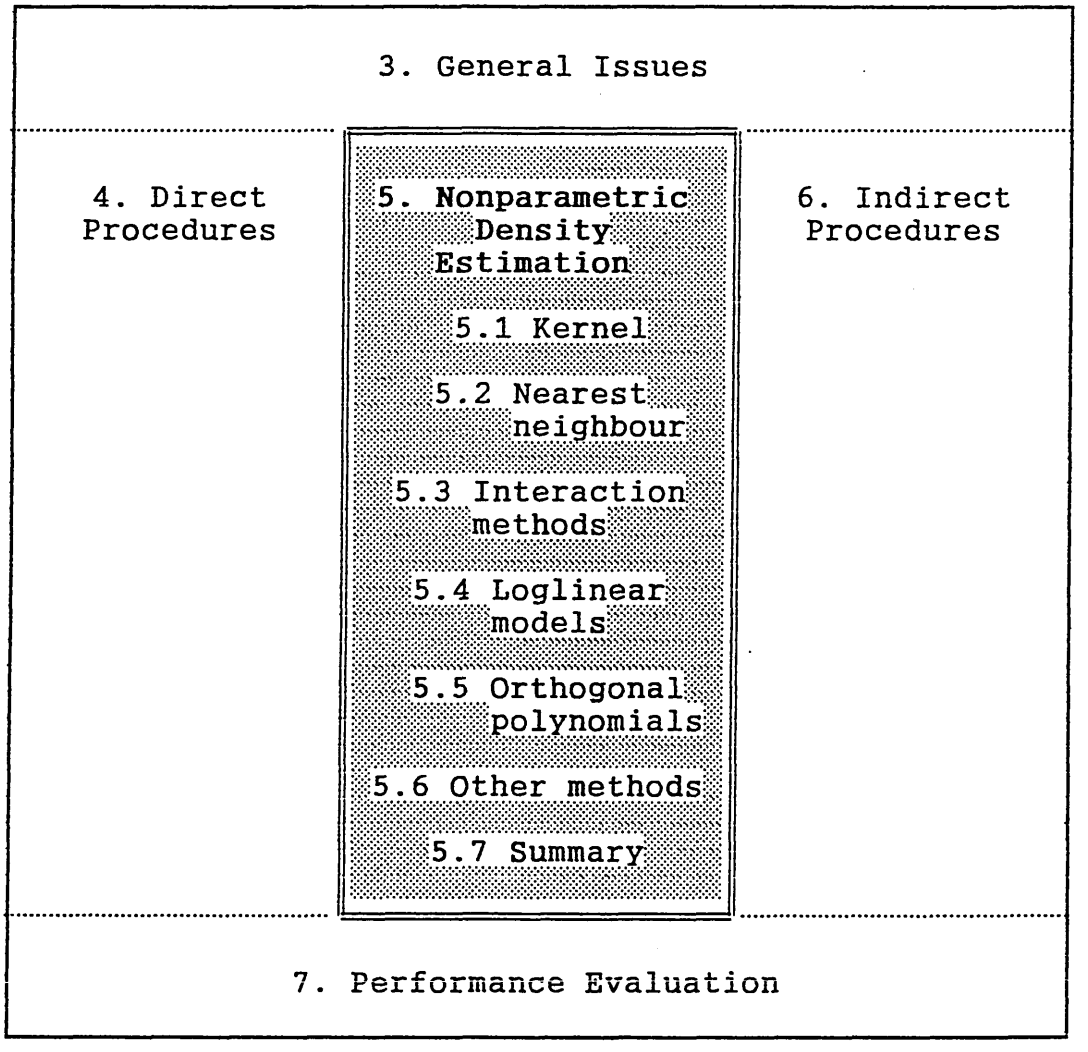
experiments. This finding, which corresponds closely to that for the nearest neighbour based discriminant procedures, is not surprising because in both types of discriminant analysis choice of appropriate smoothing parameters is not straightforward. Chapter 5 details some of the different techniques for doing this in sections 5.1 and 5.2.

4.4 Summary

For a long time direct discriminant procedures have been more popular because of their origin in classical statistics. One clear advantage is that they provide and make use of posteriors and thus readily allow the direct construction of performance measures reflecting reliability of the discriminant rule's allocations. Direct procedures also offer a wide class of density estimation techniques ranging from *parametric* through *semiparametric* to *nonparametric*. In conjunction with appropriate data models for discrete data some of these can be highly customised providing sufficient information about the underlying distributions is available. Major contributions on direct discriminant procedures include the work by Bahadur (1961), Gilchrist (1968), Cox (1972), Anderson (1975) and Goldstein & Dillon (1978).

I: INTRODUCTION

II: REVIEW



III: METHOD

IV: RESULTS

V: DISCUSSION

When looking for direct discriminant procedures, reliable estimates of the group conditional densities are required in order to compute posterior probabilities of group membership. In the discrete data situation when, as frequently is the case, no parametric data model is available, other methods of density estimation become vitally important. Some of these less common techniques are reviewed in this chapter.

Density estimation techniques for discrete data distributions[7] may be divided (figure 5-1) into *nonparametric* techniques such as kernel and nearest neighbour methods (sections 5.1 and 5.2), *parametric* techniques such as interaction models, loglinear models and models based on orthogonal polynomials (sections 5.3, 5.4 and 5.5) and a third group consisting of mixtures of parametric and nonparametric techniques (section 5.6).

class	examples
nonparametric	kernel nearest neighbour
parametric	interactive loglinear orthogonal polynomial
other	latent class smoothing techniques parametric/nonparametric combinations penalised max. likelihood

Figure 5-1: Discrete density estimation

[7] The problem of estimation of group membership relates to the estimation of prior probabilities. Due to the separate sampling scheme adopted for reasons given in chapter 10.3 this feature is therefore excluded.

The general theory of nonparametric density estimation has been reviewed by Rosenblatt (1971), Wegman (1972a, 1972b), Tarter and Kronmal (1976) and Tapia and Thompson (1978). Applications to statistical classification in particular have been reviewed by Cover and Wagner (1976).

Applications of nearest neighbour methods to problems of discriminant analysis with continuous data are traced back to Fix and Hodges (1951). Descriptions of early work are found in Loftsgarden and Quesenberry (1965) and Cover and Hart (1967). First applications of nearest neighbour methods to discrimination between two multivariate binary populations are reported by Hills (1967). These were later extended by Hall (1981a & 1981b). Kernel density estimation methods for continuous densities date back to Rosenblatt (1956), Parzen (1962) and Cacoullos (1966) and were first applied to multivariate binary discriminant analysis by Aitchison and Aitken (1976). Missing value estimation (Titterington et al, 1981), kernel based estimates for categorical data (Titterington, 1980) and estimation of smoothing parameters (Wang and von Ryzin, 1981; Hall, 1981b) laid the foundation for kernel methods in discrete discriminant analysis.

When prior knowledge about the data structure is available prior assumptions are usually in terms of parameters for underlying statistical distributions. Densities may be estimated parametrically. Frequently information is expressed as a degree of interaction present in the data. Such models are known as *interaction* models and have been investigated by Bahadur (1961a), Lancaster (1969), Moore II (1973), Victor et al (1974), Zentgraf (1975) and Trampisch (1976). Others are based on the loglinear model. Birch (1963) and Nelder and Wedderburn (1972) report some related work in this field. Finally some models are specified in terms of the order of a *Fourier series* or other *polynomial expansions* corresponding to the degree of interaction assumed present in the data. Examples of this approach are reported by Čencov (1962), Cornfield (1962), Schwartz

(1967), Specht (1967), Kronmal and Tarter (1968), Tarter and Kronmal (1970), Martin and Bradley (1972), Ott and Kronmal (1976) and Goldstein and Dillon (1978).

5.1 Kernel methods

The method of using kernels in density estimation is included as many procedures for discrete discriminant analysis utilise it (Aitchison and Aitken, 1976; Titterington et al, 1981; Hall, 1981). Of the methods to estimate probability densities of unknown functional form, the most used is the histogram. In Rosenblatt's (1956) basic paper the bias and asymptotic mean square error of a class of nonparametric estimators were derived for continuous densities. These density estimators were shown to be maximum likelihood estimators. In the 1956 paper Rosenblatt extended the histogram estimator of a probability density. Whittle (1958) proposed a smoothing method. Parzen (1962) extended Rosenblatt's work to the study of a variety of weighting functions, $K(y)$, and showed that kernel estimators are asymptotically unbiased and consistent. Cacoullos (1966) derived the estimation of a multivariate density by means of kernels.

Suppose that one wishes to estimate the density function $f(\mathbf{x})$ of an unknown distribution over a sample space S on the basis of a set $\{\mathbf{x}\}$ of n independent observations $\mathbf{x}_1,\ldots,\mathbf{x}_n$, each from this distribution. Let $K = K(\mathbf{x}|\mathbf{y},\lambda)$ denote a class of density functions on S, with mode at $\mathbf{y}$ and with λ denoting a *spread* or *concentration* parameter. A kernel estimator may be regarded as a weighted average over the empirical distribution function. A popular version of the kernel method is to take a simple mixture density of the kernel density functions such as the average density function

$$p(\mathbf{x}|\{\mathbf{x}\},\lambda) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{x}|\mathbf{y}_i,\lambda). \qquad (5.1\text{-}1)$$

as an estimate $p(\mathbf{x}|\{\mathbf{x}\},\lambda)$ of the density $f(\mathbf{x})$. For example, for $S = R^q$ the kernel density function adopted after preliminary scaling of the data is often the circular or spherical normal (Habbema et al, 1974)

$$K(\mathbf{x}|\mathbf{y},\lambda) = \frac{1}{(2\pi\lambda)^{q/2}}\, e^{-\frac{1}{2}\frac{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}{\lambda}}. \qquad (5.1\text{-}2)$$

The variance parameter, λ, has to be chosen carefully. If λ is too small the resulting density function becomes peaked and sample dependent. If λ is too large the resulting density function becomes uniform. Aitchison and Aitken (1976) applied the method to multivariate binary data. Here $S = B^q$, q-dimensional binary space with $B = \{0,1\}$. The convenient location-scale property assigned to K for density estimation in R^q, namely that $K(\mathbf{x}|\mathbf{y},\lambda)$ is expressible in the form $\lambda^{-1}H\{(\mathbf{x}-\mathbf{y})/\lambda\}$, is not available for B^q. For binary data Aitchison and Aitken defined as a counterpart of the spherical normal kernel (5.1-2) the following

$$K(\mathbf{x}|\mathbf{y},\lambda) = \lambda^{q-d(\mathbf{y},\mathbf{x})}(1-\lambda)^{d(\mathbf{y},\mathbf{x})} \qquad (5.1\text{-}3)$$

where $\frac{1}{2} \le \lambda \le 1$; $\mathbf{x},\mathbf{y} \in B^q$ and $d(\mathbf{y},\mathbf{x}) = (\mathbf{y}-\mathbf{x})'(\mathbf{y}-\mathbf{x})$. The dissimilarity coefficient $d(\mathbf{y},\mathbf{x})$, while expressible in terms of a squared euclidean distance, is simply the number of disagreements in corresponding components of $\mathbf{y}$ and $\mathbf{x}$. Since for all $\mathbf{y},\mathbf{x} \in B^q$,

$$K\left(\mathbf{x}|\mathbf{y},\frac{1}{2}\right) = \left(\frac{1}{2}\right)^q \qquad (5.1\text{-}4)$$

$$K(\mathbf{x}|\mathbf{y},1) = \begin{cases} 1 & (\mathbf{x} = \mathbf{y}) \\ 0 & (\mathbf{x} \neq \mathbf{y}) \end{cases} \qquad (5.1\text{-}5)$$

the end points of the range of the smoothing parameter λ provide two extreme forms of estimation. Thus $\lambda = \frac{1}{2}$ gives the uniform distribution over B^q whatever the data, and $\lambda = 1$ estimates the densities simply by utilising corresponding relative frequencies. This is equivalent to estimated cell frequencies in the multinomial model. Methods for choosing optimal values for the smoothing parameter λ generally involve maximising a likelihood function of the data $\mathbf{x}$ with respect to λ for a given kernel function $K(\mathbf{x}|\mathbf{y},\lambda)$. This approach is used by Habemma et al (1974), Aitchison and Aitken (1976), Titterington (1977), Titterington (1980), Hall (1981) and Wang and van Ryzin (1981) among others. Multi-category variables and incomplete data can generally be coped with in all types of kernel. Another advantage of kernel estimators lies in the small number of parameters required. The estimator (5.1-2) of Aitchison and Aitken (1976) requires estimation of no more than one λ_i per class. Titterington (1980) points out that in its most general form (5.1-3) takes the shape

$$K(\mathbf{x}|\mathbf{y},\lambda) = \prod_{i=1}^{q} \lambda_i^{1-d_i(y,x)}(1-\lambda_i)^{d_i(y,x)} \qquad (5.1\text{-}6)$$

where $d_i(\mathbf{y},\mathbf{x}) = 0$ if the i^{th} components of $\mathbf{y}$ and $\mathbf{x}$ are the same and 1 otherwise and λ has components $\lambda_1,\ldots,\lambda_q$. The number of parameters in (5.1-6) can be reduced to 1 per group by taking all the λ_i's equal. It is more helpful, however, to retain different values because of the possibly different variances of the components of $\mathbf{x}$. A variety of kernel functions are supplied by statistical software packages. The *SAS* package for instance offers a choice among *4* different types: *normal*, *Epanechnikov*, *biweight* and *triweight*. The manual gives no suggestion about which type of function to select and is also rather vague regarding the choice of the smoothing parameter. It suggests that "for a fixed kernel shape, one way to choose the smoothing parameter, λ, is to plot estimated densities with different values of r and to choose the estimate *that is most in accordance* with the prior information about the density."

<Italics not in original>. Whilst apparently offering very pragmatic advice this quote highlights again the central problem of appropriate choice of the smoothing parameter λ[8]. From the reports in the literature choice of λ does appear to be far more important than the type of kernel function chosen for a given application.

5.2 Nearest neighbour methods

Nearest neighbour techniques for continuous data were described by Fix and Hodges (1951) and Loftsgaarden and Quesenberry (1965). Hills (1967) defines a nearest neighbour procedure for binary data in the two population situation where the distance is defined in terms of number of disagreements between components of q-variate binary vectors. Consider an object with $\mathbf{X}=\mathbf{x}_0$ which is allocated to Π_1 or Π_2. The likelihood ratio would estimate $LR(\mathbf{x}_0)$ by the ratio $(r_1/n_1)/(r_2/n_2)$ where r_i objects of the sample have $\mathbf{X}=\mathbf{x}_0$ in Π_i. A *1*-nearest neighbour rule also considers those r_i' near neighbours in the sample from Π_i whose $\mathbf{x}$ values differ from $\mathbf{x}_0$ in respect of only one variable. The likelihood ratio at $\mathbf{x}_0$ is estimated by

$$\left\{\frac{r_1+r_1'}{n_1}\right\} \Big/ \left\{\frac{r_2+r_2'}{n_2}\right\} . \qquad (5.2\text{-}1)$$

Similarly, for a *2*-nearest neighbour rule $LR(\mathbf{x}_0)$ is estimated by

$$\left\{\frac{r_1+r_1'+r''}{n_1}\right\} \Big/ \left\{\frac{r_2+r_2'+r_2''}{n_2}\right\} \qquad (5.2\text{-}2)$$

---

[8] This problem is not particular to kernel density estimation but applies to non-parametric density estimation in general (see also section 5.2).

where r_i'' of the sample members from Π_i have $\mathbf{x}$ values differing from $\mathbf{x}_0$ in respect of two variables. A nearest neighbour rule of any order up to q may be defined in the same fashion. The *0*-nearest neighbour rule corresponds to the maximum likelihood rule in the multinomial model. The advantage of the nearest neighbour rule over the maximum likelihood rule is that the estimate of $LR(\mathbf{x}_0)$ will be less subject to sampling variation. On the other hand, it is not necessarily consistent. This is illustrated clearly when the $X_1,\ldots,X_q$ are independent in both Π_1 and Π_2. Then the average of the loglikelihood ratios over the point $\mathbf{x}_0$ and its q near neighbours of order *1* is

$$\frac{qL(\mathbf{x}_0) + L(\mathbf{y})}{q + 1}\ , \qquad (5.2\text{-}3)$$

where $\mathbf{y}=(1-x_1,1-x_2,\ldots,1-x_q)'$, the mirror image of $\mathbf{x}_0$ as $\mathbf{x}$ is multivariate binary. By restricting the variables on which near neighbours are allowed to differ to unimportant ones it may be ensured that $L(\mathbf{y})$ will not differ too greatly from $L(\mathbf{x}_0)$ and hence, especially for large q, the weighted average of $L(\mathbf{x}_0)$ and $L(\mathbf{y})$ will be close to $L(\mathbf{x}_0)$.

Hall (1981) constructs nearest neighbour estimators based on optimal linear combinations in the sense of minimising mean squared error. Given two symptom combinations $\mathbf{x},\mathbf{y} \in \{0,1\}^q \equiv B^q$, their *distance* apart or number of disagreements is given by $d(\mathbf{x},\mathbf{y})=(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})$. If S is a sample of n symptom combinations from a certain combination $\mathbf{b}\in B^q$ is given by $N_j(\mathbf{b})$, the number of $\mathbf{x}\in S$ with $d(\mathbf{b},\mathbf{x})=j$. Hills (1967) proposed near neighbour estimates of order *1* ($0 \le 1 \le q-1$) of the relative probabilities of observing the combination $\mathbf{b}$:

$$\hat{p}_H(\mathbf{b}) = n^{-1}\sum_{j=0}^{1} N_j(\mathbf{b}). \qquad (5.2\text{-}4)$$

The estimators suggested by Hall (1981) are weighted equivalents of the form:

$$\hat{p}_o(b) = n^{-1} \sum_{j=0}^{1} \omega_j N_j(b), \qquad (5.2\text{-}5)$$

where the weights ω_j are chosen to minimise $\Delta(\omega_0,\ldots,\omega_1) = \Sigma_b E\{\hat{p}_o(b)-p_o(b)\}^2$. Hall finds that the probabilities calculated using weighted near neighbour estimators are similar to those computed from the kernel technique of Aitchison and Aitken (1976) but both quite different from those obtained by using Hills' (1967) near neighbour estimator when applied to the *keratoconjunctivitis sicca* data of Anderson et al (1972).

The *k*-nearest neighbour method as described in the *SAS* manual (1986) for the version 6 edition uses the Mahalanobis distances based on covariance matrices as a metric. The number, *k*, of training set points for each observation **x** is fixed. The method finds the radius $r_k(\mathbf{x})$ which is the distance from **x** to the k^{th} nearest training set point in the metric $\mathbf{V}_i^{-1}$. Considering a closed ellipsoid centered at **x** and bounded by $\{\mathbf{z}|(\mathbf{z}-\mathbf{x})'\mathbf{V}_i^{-1}(\mathbf{z}-\mathbf{x}) = r_k^2(\mathbf{x})\}$ the nearest neighbour method is equivalent to the uniform kernel method with a location dependent radius $r_k(\mathbf{x})$. Using the *k*-nearest neighbour rule the *k* smallest distances are saved. Of these *k* distances let k_i represent the number of distances that are associated with group *i*. Then as in the uniform kernel method the estimated group *i* density at **x** is $\hat{f}_i(\mathbf{x}) = k_i/(n_i v_k(\mathbf{x}))$ where $v_k(\mathbf{x})$ is the volume of the ellipsoid bounded by $\{\mathbf{z}|(\mathbf{z}-\mathbf{x})'\mathbf{V}_i^{-1}(\mathbf{z}-\mathbf{x}) = r_k^2(\mathbf{x})\}$. Since the pooled within group covariance matrix is used to calculate the distances used in the nearest neighbour method, the volume $v_k(\mathbf{x})$ is a constant independent of group membership. The *SAS* manual is again rather unspecific in advice regarding the choice of *k*: "A practical approach is to try several values of the smoothing parameters within the context of the particular application and to choose the one which gives the most satisfactory results." Hand (1982) has indeed found that the choice of *k* is usually relatively uncritical.

On the basis of a paper by Lazarsfeld (1956) who obtained similar representations, Bahadur (1961a) derived a representation of the joint distribution of responses to n dichotomous items and in a further paper (1961b) developed a classification procedure based on his model. For $\mathbf{x}=(x_1,\ldots,x_q)'$ with $\sum_{\mathbf{x}\in B^q} f(\mathbf{x})=1$ let the means

$$\mu_j = E_f(x_j), \qquad 0 < \mu_j < 1;\ j=1,\ldots,q \qquad (5.3\text{-}1)$$

where E_f denotes the expected value when f obtains. Next, setting

$$Z_j = \frac{x_j-\mu_j}{\sqrt{\mu_j(1-\mu_j)}} \qquad (5.3\text{-}2)$$

the correlations

$$\begin{aligned} \rho_{jk} &= E_f(Z_jZ_k) \quad , j < l; \qquad (5.3\text{-}3)\\ \rho_{jkl} &= E_f(Z_jZ_kZ_l) \quad , j < l < k;\\ &\ldots\\ &\ldots\\ \rho_{12\ldots q} &= E_f(Z_1Z_2\ldots Z_q) \end{aligned}$$

are defined. There are thus C_2^q second order correlations r_{jk}, C_3^q third order correlations r_{jkl}, and so on up to *1* q^{th} order correlation $r_{12\ldots q}$.

Bahadur shows (1961a) that the joint probability distribution of $\mathbf{x}$ may be written as

$$f(\mathbf{x}) = \prod_{j=1}^{q} \mu_j^{x_j}(1-\mu_j)^{1-x_j} \times \qquad (5.3\text{-}4)$$

$$\{\ 1 + \sum_{j<k} \rho_{jk}z_jz_k + \sum_{j<k<} \rho_{jkl}z_jz_kz_l + \ldots + \rho_{12\ldots q}z_1z_2\ldots z_q\ \}.$$

Note that the first term in (5.3-4) corresponds to the independence model. In all there are $\sum_{v=2}^{q} C_v^q + q = 2^q - 1$ parameters which equal the number in the full multinomial representation. The Bahadur representation is a direct expression in terms of means and correlations. Applied to the case of two populations the means μ_1, μ_2 and correlations ρ_1, ρ_2 for $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are estimated separately and a suitable expression for the likelihood ratio based on (5.3-4) may then be used for the derivation of allocation rules. Solomon (1961) first applied the Bahadur representation to a set of multivariate binary data with $q = 4$ on attitudes to science of two groups of high school seniors with different intelligence quotients.

The interaction model due to Lancaster (1969) may be written as in Zentgraf (1975) as

$$p_{j_1..j_q} = \sum_{v=2}^{s} (-1)^{s-v} \binom{q-v-1}{s-v} \sum_{C_q^v} p_{j_k..j_{k_v}}^{(k_1..k_v)} \; p_{j_m}^{(m_1)} .. p_{j_{m-v}}^{(m_{q-v})}$$

$$+ (-1)^s \left[\binom{q-1}{s} - q \binom{q-2}{s-1} \right] \prod_{m=1}^{q} p_{j_m}^{(m)} \qquad (5.3\text{-}5)$$

where C_q^v stands for the set of all combinations of v elements out of $\{1,\ldots,q\}$. Here $p_{j_{1\ldots}j_q}$ is a cell probability, i.e. the joint probability for the variables X_m to have outcomes j_m, $m=1,\ldots,q$. Further, $p_j^{(m)}$ stands for the probability of the j^{th} outcome of X_m, $p_{jk}^{(nm)}$ for the probability of the joint occurrence of the j^{th} outcome of X_n and the k^{th} occurrence of X_m, etc. This corresponds to the index-dot notation, with the modification that only those indices which are not substituted by a dot are given and their position is marked by a superscript. Zentgraf gives further expressions for (5.3-5) when higher order interactions vanish, in particular for $s=2$.

The main advantage of (5.3-5) is that it is possible to calculate the cell probabilities explicitly by use of

marginal probabilities of first up to s^{th} order if disappearance of interactions higher than s^{th} order is assumed. The relative frequencies may be used as simple estimates of the marginal probabilities. These estimates are stable for sufficiently small s and hence also for the estimates of the cell probabilities. For dichotomous variables (5.3-5) the densities may be obtained by another approach, using the symmetric parameters introduced by Lazarsfeld (1961) or equivalently using the n^{th} order correlation parameters introduced by Bahadur (1961a). The same result will be achieved with these representations since the vanishing of Lancaster's s^{th} order interactions, the disappearance of Lazarsfeld's $(s+1)^{th}$ order symmetric parameters and the vanishing of Bahadur's $(s+1)^{th}$ order correlations are all equivalent. Goldstein and Dillon (1978) also showed that Bahadur models for multivariate binary data are contained in Lancaster models for discrete data.

5.4 Loglinear models

A final set of models that bears on the topic of density estimation of discrete data is that of loglinear models as first suggested for three dimensions by Birch (1963). This was extended by Grizzle, Starmer and Koch (1969) and later by Nelder and Wedderburn (1972) to the theory of generalised linear models. Of special interest are models that express the logarithm of the state probabilities in terms of a linear combination of main effects and interactions. Suppose that

$$\log f(\mathbf{x}) = \alpha + \sum_{j=1}^{q} (-1)^{x_j} \alpha_j + \sum_{j<k} (-1)^{x_j+x_k} \alpha_{jk} + \ldots$$
$$\ldots + (-1)^{x_1+x_2+\ldots+x_q} \alpha_{12\ldots q} \qquad (5.4\text{-}1)$$

where α is an overall effect, α_j is the main effect due to X_j and α_{jk} is the respective interaction effect between X_j and X_k and so on. The discrete density is specified by estimating respective main effect and interactive terms in

(5.4-1). The vector x includes the factor containing group membership, x_1, say.

5.5 Procedures based on orthogonal polynomials

Martin and Bradley (1972) proposed for $X \in B^q$ the estimator

$$f_i(x) = f(x) \left[1 + h(a^{(i)},x)\right] \qquad (5.5\text{-}1)$$

where $h(a^{(i)},x)$ is a polynomial in the elements of x and the coefficients $a^{(i)}$ are specific to Π_i ($i=1,\ldots,g$) and $f(x)$ is the weighted sum $\sum_i p_i f_i(x)$. The function $h(a^{(i)},x)$ is expressed in terms of orthogonal polynomials $\phi_g(x)$ where

$$\phi_0(x) = 1, \ \Phi_j(x) = 2x_j - 1, \ j=1,\ldots,q \qquad (5.5\text{-}2)$$

$$\phi_g(x) = \prod_{j=1}^{k} \phi_{\gamma_j}(x) \quad \gamma=(\gamma_1,\ldots,\gamma_k)', \ \gamma_1<\gamma_2<\ldots<\gamma_k,$$
$$k = 2,\ldots,q, \quad \gamma_j \in \{1,\ldots,q\}.$$

The complete set of 2^q values of γ is denoted by Γ_q indicating all polynomial terms up to and including order q. The orthogonal property follows from

$$\sum_{x \in B^q} {}_g(x)\phi_0(x) = 2^q \Delta(\gamma,\delta) \quad \gamma,\delta \in \Gamma_q, \qquad (5.5\text{-}3)$$

where $\Delta(\gamma,\delta)=1,0$ as $\gamma=,\neq\delta$. As the set of 2^q polynomials $\phi_g(x)$, $\gamma \in \Gamma_q$, forms a basis for the set of all real valued functions defined on B^q it follows that for any set of probability functions $f_i(x)$ ($i=1,\ldots,g$) one may write

$$h(a^{(i)},x) = \sum_{\gamma \in \Gamma_q} {}_g{}^{(i)}\phi_g(x). \qquad (5.5\text{-}4)$$

Equations (5.5-1) and (5.5-4) show that

$$h(a^{(i)},x) = \frac{f_1(x) - f(x)}{f(x)} \qquad (5.5\text{-}5)$$

and

$$a_g^{(i)} = 2^{-q} \sum_{x \in B^q} \phi_g(x) \frac{f_1(x) - f(x)}{f(x)} \qquad (5.5\text{-}6)$$

for all $i=(1,\ldots,g)$ and $\gamma \in \Gamma_q$ provided $f(x) \neq 0$. In the case of independent random samples available from Π_i, maximum likelihood estimates for $f_i(x)$ are $\hat{f}_i(x)=n_i(x)/n_i$ where $n_i(x)$ is the frequency of observations with state x. The estimators are then

$$\hat{f}(x) = \textstyle\sum_i \pi_i \hat{f}_i(x) \text{ and} \qquad (5.5\text{-}7)$$

$$\hat{a}_g^{(i)} = 2^{-q} \sum_{x \in B^q} \phi_g(x) \frac{\hat{f}_i(x) - \hat{f}(x)}{\hat{f}(x)} \qquad (5.5\text{-}8)$$

again provided $\hat{f}_i(x) \neq 0$. When all 2^q parameters are estimated the Martin and Bradley model is equivalent to the full multinomial rule. Potentially useful models are chosen by deletion of selected parameters in the expansion of $h(a^{(i)},x)$. The authors suggest fitting a reduced model of the form

$$f_i(x) = f(x) \left[1 + h_s(a^{(i)},x)\right] \qquad (5.5\text{-}9)$$

where s denotes a particular order of subset of polynomials, usually corresponding to main effects and low-order interactions. If Γ_s represents the set of polynomials of maximum order s then

$$h_s(a^{(i)},x) = \sum_{\gamma \in \Gamma_s} g^{(i)} \phi_g(x). \qquad (5.5\text{-}10)$$

Iterative methods are then employed to obtain maximum likelihood estimates for the $f_i(x)$. Martin and Bradley apply their model to the Solomon (1961) data on attitudes

towards science and to *16* state multivariate binary hypoxic trauma data, known as the Martin-Lamper data. A χ^2 test of fit yielded the value *16.2* with *11* degrees of freedom for the Solomon data which is not significant at the $\alpha=0.05$ level indicating a moderately good fit. The correspondence between multinomial estimates for cell frequencies and first order model estimates is also reasonably good also for the Martin-Lamper data.

Kronmal and Tarter (1968) devised a method of density estimation using Fourier series. Ott and Kronmal (1976) later estimated the multivariate binary density by an orthogonal expansion of the density in terms of discrete Fourier series and derived four variants for this method. Together with three standard methods, the independence model, the logistic model and the full multinomial the methods are compared on *11 6*-variable datasets generated by Monte Carlo sampling with differing degrees of interactive structure. The results of the sampling experiments indicate that the independence method does best in general even when applied to datasets with high degrees of interaction. The authors note a tendency for the independence model to perform worse with increases in sample size to around *1000* cases.

5.6 Other methods of density estimation

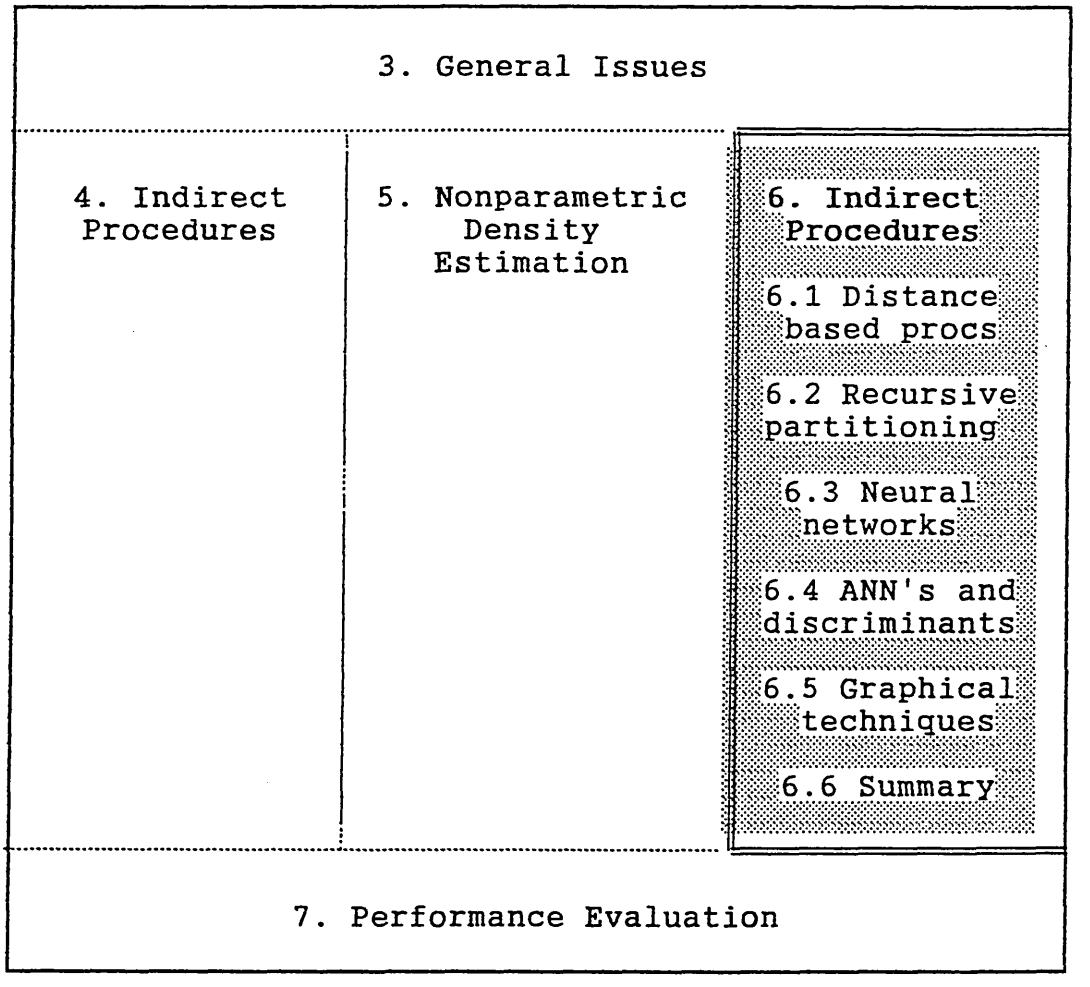
Other approaches to density estimation for discrete data include Lazarsfeld's (1960) latent class model, Whittles's (1958) smoothed estimates determined by Bayesian methods and Dickey's (1968) data-analytic approach to smoothed density estimates based on natural stationarity assumptions. A comprehensive review of adaptive robust procedures is found in Hogg (1974). Olkin and Spiegelman (1987) adopt a different approach via convex combinations of parametric and nonparametric density estimates. A recent paper by Granville and Rasson (1995) discusses density estimation via *penalised maximum likelihood.*

5.7 Summary

The variety of density estimation techniques for discrete data warrants treatment in a separate chapter. While there exists a breadth of *kernel* based methods, also highly adapted to discrete data situations, it appears that the *degree of smoothing* used is more important than the type of kernel chosen. A similar finding also holds for the *nearest neighbour* techniques. For a considerable time statistical models (Bahadur and Lancaster) have also existed for describing multivariate discrete distributions that allow direct modelling in terms of means and interactions. Other forms of density estimation less used in discriminant analysis applications include *loglinear models* and *orthogonal polynomials.* Major work on non-parametric density estimation with relevance to discriminant analysis for discrete data, includes the contributions by Fix & Hodges (1951), Bahadur (1961), Rosenblatt (1971), Aitchison & Aitken (1976), and Titterington et al (1981).

I: INTRODUCTION

II: REVIEW



III: METHOD

IV: RESULTS

V: DISCUSSION

Chapter 6 - Indirect discriminant procedures

Indirect procedures provide allocation rules that are not based on posterior probabilities of population membership. For this reason density estimation is not a feature of this growing class of discriminant procedures. Their advantage clearly lies in the distribution free property. Indirect procedures tend to be computing intensive and are thus generally more recent. The subclass of indirect procedures is reviewed in this chapter in 5 sections. Section 6.1 treats *distance based procedures*, section 6.2 *recursive partitioning procedures*, section 6.3 reviews the field of *artificial neural networks* (*ANN*), section 6.4 discusses suitability of *ANN's* for problems of discriminant analysis from a theoretical point of view and section 6.5 reviews procedures based on *graphical techniques*.

6.1 Distance based procedures

Distance methods are nonparametric and depend on the suitable definition of a distance measure where the metric may be the *identity metric* giving the euclidean distance or scaled by a variance-covariance matrix or defined in terms of functions of individual state probabilities. The *centroid method* employed by Moore (1982) or Pridmore (1985) uses the euclidean distance. Goldstein and Dillon (1978) use a distance measure based on Matusita's (1955) definition of distance. A comparison between these two methods will be given here in some detail, as the distributional distance method, in particular, has not appeared very much in the literature. Yet this may constitute an attractive nonparametric approach to discrimination.

The centroid method is probably the simplest most general rule and essentially depends on computing distances, d_{ij}^2 with $i=1,2,\ldots,g$ between individual observations, X_j, and the centres of gravity of the g populations. X_j is then

allocated to the population Π_i giving rise to the least distance, d_{ij}^2. In the case of *2* populations and corresponding *q*-variate mean vectors μ_1 and μ_2 the allocation of a new observation X_j with unknown group membership is according to the rule: Allocate X_j to the population Π_1 if

$$d_{1j}^2 < d_{2j}^2, \qquad \text{or}$$

$$\sum_{k=1}^{q} \left(X_{ijk}-\mu_{1k}\right)^2 < \sum_{k=1}^{q} \left(X_{ijk}-\mu_{2k}\right)^2. \qquad (6.1\text{-}1)$$

Moore (1982) suggests utilising prior probabilities by weighting the distances by the inverse square of the priors:

$$d_{ij}^{2*} = \frac{1}{\pi_j^2} d_{ij}^2 \qquad (6.1\text{-}2)$$

where d_{ij}^{2*} is the new distance. This approach was not adopted but instead an empirical algorithm minimising the misallocation error ε was developed. Goldstein and Dillon (1978) proposed a classification rule based on the distributional distance between populations Π_1 and Π_2 derived by Matusita (1955). This distance is given by

$$d^{\text{Matusita}} = \|\mathbf{p}_1-\mathbf{p}_2\|^2 = \sum_{j=1}^{s} \left(\sqrt{p_{1j}} - \sqrt{p_{2j}}\right)^2 \qquad (6.1\text{-}3)$$

where $\|.\|$ is the *vector length* operator[9] and $\mathbf{p}_1$ and $\mathbf{p}_2$ are as defined below. Note that d^{Matusita} may be the same for different levels of state probabilities, p_{ij}. For the two population situation, the vectors $\mathbf{p}_1=(p_{11},p_{12},\ldots,p_{1s})'$ and $\mathbf{p}_2=(p_{21},p_{22},\ldots,p_{2s})'$ define the probability distributions in Π_1 and Π_2. The number of discrete states for the two

---

[9] for q-variate $\mathbf{X}$: $\|\mathbf{X}\| = \sqrt{X_1^2+X_2^2+\ldots+X_q^2}$

populations Π_1 and Π_2 is given by s. The set of parameters $\{p_{ij}\}$ with $j=1,\ldots,s$ are the probabilities observed for multinomial random variables $X^{(s)}$ written in the state matrix notation[10]. Let $\mathbf{p}_i^{+1}$ be the vector of state probabilities in population Π_i after inclusion of one further multinomial observation $X_v^{(s)}$. Then $X_v^{(s)}$ designates the multinomial observation of the v^{th} state. Given such a new observation the allocation rule is

$$\text{allocate } X_v^{(s)} \begin{cases} \text{to } \Pi_1 \text{ if } \|\mathbf{p}_1^{+1}-\mathbf{p}_2\| > \|\mathbf{p}_1-\mathbf{p}_2^{+1}\| \\ \text{to } \Pi_2 \text{ if } \|\mathbf{p}_1^{+1}-\mathbf{p}_2\| < \|\mathbf{p}_1-\mathbf{p}_2^{+1}\| \\ \text{randomly if } \|\mathbf{p}_1^{+1}-\mathbf{p}_2\| = \|\mathbf{p}_1-\mathbf{p}_2^{+1}\| \end{cases} . \qquad (6.1\text{-}4)$$

The first of the inequalities in (6.1-4) states that a new observation $X_v^{(s)}$ is allocated to the population that will result in a greater interpopulation distance, d^{Matusita}, when inserted into expression 6.1-3. Upon expansion of the length operator in 6.1-4 the first inequality becomes

$$\sum_{j=1}^{s}\left(\sqrt{\frac{n_{1j}^{+1}}{n_1+1}}-\sqrt{\frac{n_{2j}}{n_2}}\right)^2 > \sum_{j=1}^{s}\left(\sqrt{\frac{n_{1j}}{n_1}}-\sqrt{\frac{n_{2j}^{+1}}{n_2+1}}\right)^2 . \qquad (6.1\text{-}5)$$

The term n_{ij}^{+1} is defined for $i=1,\ldots,g$ by

$$n_{ij}^{+1} = \begin{cases} n_{ij}+1 & \text{for } v = j \\ n_{ij} & \text{for } v \neq j \end{cases} . \qquad (6.1\text{-}6)$$

By summing separately for the two conditions given in expression 6.1-6 the inequality 6.1-5 may be seen to be

---

10 i.e. $X^{(2)}$ is multinomial with S states.

$$\left(\sqrt{\frac{n_{1v}+1}{n_1+1}} - \sqrt{\frac{n_{2v}}{n_2}}\right)^2 + \sum_{j\neq v}\left(\sqrt{\frac{n_{1j}}{n_1+1}} - \sqrt{\frac{n_{2j}}{n_2}}\right)^2$$

$$> \qquad (6.1\text{-}7)$$

$$\left(\sqrt{\frac{n_{1v}}{n_1}} - \sqrt{\frac{n_{2v}+1}{n_2+1}}\right)^2 + \sum_{j\neq v}\left(\sqrt{\frac{n_{1j}}{n_1}} - \sqrt{\frac{n_{2j}}{n_2+1}}\right)^2$$

$$\Leftrightarrow$$

$$\frac{n_{1v}+1}{n_1+1} - 2\sqrt{\frac{(n_{1v}+1)n_{2v}}{(n_1+1)n_2}} + \frac{n_{2v}}{n_2} + \sum_{j\neq v}\frac{n_{1j}}{n_1+1} - 2\sum_{j\neq v}\sqrt{\frac{n_1 n_2}{(n_1+1)n_2}} + \sum_{j\neq v}\frac{n_{2j}}{n_2}$$

$$> \qquad (6.1\text{-}8)$$

$$\frac{n_{1v}}{n_1} - 2\sqrt{\frac{(n_{2v}+1)n_{1v}}{(n_2+1)n_1}} + \frac{n_{2v}+1}{n_2+1} + \sum_{j\neq v}\frac{n_{2j}}{n_2+1} - 2\sum_{j\neq v}\sqrt{\frac{n_1 n_2}{(n_2+1)n_1}} + \sum_{j\neq v}\frac{n_{1j}}{n_1}$$

$$\Leftrightarrow$$

$$\frac{n_{1v}+1}{n_1+1} + \frac{n_{2v}}{n_2} + \sum_{j\neq v}\frac{n_{1j}}{n_1+1} + \sum_{j\neq v}\frac{n_{2j}}{n_2} - 2\left(\sqrt{\frac{(n_{1v}+1)n_{2v}}{(n_1+1)n_2}} + \sum_{j\neq v}\sqrt{\frac{n_1 n_2}{(n_1+1)n_2}}\right)$$

$$> \qquad (6.1\text{-}9)$$

$$\frac{n_{1v}}{n_1} + \frac{n_{2v}+1}{n_2+1} + \sum_{j\neq v}\frac{n_{1j}}{n_1} + \sum_{j\neq v}\frac{n_{2j}}{n_2+1} - 2\left(\sqrt{\frac{(n_{2v}+1)n_{1v}}{(n_2+1)n_1}} + \sum_{j\neq v}\sqrt{\frac{n_1 n_2}{(n_2+1)n_1}}\right)$$

$$\Leftrightarrow$$

$$\frac{n_{1v}+1+\sum_{j\neq v} n_{1j}}{n_1+1} + \frac{n_{2v}+\sum_{j\neq v} n_{2j}}{n_2} - \frac{2}{\sqrt{n_2(n_1+1)}} \left\{ \sqrt{\frac{n_{2v}}{n_{1v}+1}} + \sum_{j\neq v} \sqrt{n_{1j}n_{2j}} \right\}$$

$$> \qquad (6.1\text{-}10)$$

$$\frac{n_{1v}+\sum_{j\neq v} n_{1j}}{n_1} + \frac{n_{2v}+1+\sum_{j\neq v} n_{2j}}{n_2} - \frac{2}{\sqrt{n_1(n_2+1)}} \left\{ \sqrt{\frac{n_{1v}}{n_{2v}+1}} + \sum_{j\neq v} \sqrt{n_{1j}n_{2j}} \right\}$$

$$\Leftrightarrow$$

$$\frac{1}{\sqrt{n_2(n_1+1)}} \left\{ \sqrt{\frac{n_{2v}}{n_{1v}+1}} + \sum_{j\neq v} \sqrt{n_{1j}n_{2j}} \right\}$$

$$< \qquad (6.1\text{-}11)$$

$$\frac{1}{\sqrt{n_1(n_2+1)}} \left\{ \sqrt{\frac{n_{1v}}{n_{2v}+1}} + \sum_{j\neq v} \sqrt{n_{1j}n_{2j}} \right\}$$

$$\Leftrightarrow$$

$$\frac{\sqrt{\frac{n_{2v}}{n_{1v}+1}} + \sum_{j\neq v} \sqrt{n_{1j}n_{2j}}}{\sqrt{\frac{n_{1v}}{n_{2v}+1}} + \sum_{j\neq v} \sqrt{n_{1j}n_{2j}}} < \sqrt{\frac{n_2(n_1+1)}{n_1(n_2+1)}} \; . \qquad (6.1\text{-}12)$$

Upon making the resubstitutions

$$c_1 = \sqrt{\frac{n_{1v}}{n_{2v}+1}} \quad , \quad c_2 = \sqrt{\frac{n_{2v}}{n_{1v}+1}}$$

$$a = \sum_{j \neq v} \sqrt{n_{1j} n_{2j}} \quad , \quad \lambda = \sqrt{\frac{n_2(n_1+1)}{n_1(n_2+1)}} \qquad (6.1\text{-}13)$$

the allocation rule 6.1-12 may be simplified to

$$\text{"allocate } X_v^{(s)} \text{ to } \Pi_1 \text{ if } \frac{c_2 + a}{c_1 + a} < \lambda \text{"} \; . \qquad (6.1\text{-}14)$$

An example for the two population situation with dichotomous data with $s=2^3=8$ discrete states is quoted from Goldstein and Dillon (1978) and given in tables 6.1-1 and 6.1-2.

		Π_1		Π_m	
	state	freq.	rel. freq.	freq.	rel. freq.
j	$(Y_1Y_2Y_3)$	(l_j)	(l_j/l)	(m_j)	(m_j/m)
1	1 1 1	13	0.520	20	0.073
2	1 1 0	5	0.200	15	0.055
3	1 0 1	0	0.000	2	0.007
4	1 0 0	1	0.040	50	0.182
5	0 1 1	3	0.120	86	0.313
6	0 1 0	0	0.000	1	0.004
7	0 0 1	1	0.040	1	0,004
8	0 0 0	2	0.080	100	0.364
		25(l)	1.000	275(m)	1.000

Table 6.1-1: Generalised distance example

Table 6.1-1 shows an 8-state example of the generalised distance rule (Goldstein and Dillon, 1978). Application of (6.1-14) yields the results shown in table 6.1-2.

state	counts		criterion	comparison	allocation
j	l_j	m_j	$\frac{c_2 + a}{c_1 + a}$	λ	to
1	*13*	*20*	*1.003*	*1.018*	Π_1
2	*5*	*15*	*1.008*	*1.018*	Π_1
3	*0*	*2*	*1.022*	*1.018*	Π_2
4	*1*	*50*	*1.045*	*1.018*	Π_2
5	*3*	*86*	*1.038*	*1.018*	Π_2
6	*0*	*1*	*1.016*	*1.018*	Π_1
7	*1*	*1*	*1.000*	*1.018*	Π_1
8	*2*	*100*	*1.049*	*1.018*	Π_2

Table 6.1-2: Computing the generalised distance

Note the results for the 6^{th} state $(0,1,0)$, where $l_j=0$ but allocation is to Π_1 ($1.016 < 1.018 \Rightarrow \Pi_1$). For the two non-parametric methods, however, no direct expression for the optimal cutoff point, k_{opt}, is easily derived as the distribution of the quantities (criterion values)

$$d_{1j}^2 - d_{2j}^2 = \sum_{k=1}^{q} \left[X_{jk} - \mu_{1k}\right]^2 - \sum_{k=1}^{q} \left[X_{jk} - \mu_{2k}\right]^2 \qquad (6.1\text{-}15)$$

and

$$\log_e(R_v) = \log_e\left[\frac{c_2 + a}{c_1 + a} \cdot \frac{1}{\lambda}\right] \qquad (6.1\text{-}16)$$

for the centroid method and for the distance method, in particular, are difficult to derive. As mentioned above, Moore (1982) does give a suggestion as to how prior probabilities may be incorporated via the inverse square of the distance for the centroid method, but no indication is given for the distance method. Without going into the details of examining the adequacy of the suggestion by Moore, an alternative numerical solution was employed by Lack (1987). This approach has the added advantage of full generalisability and is described in the following.

In each of the two methods, centroid and distributional distance, the expression for the total misallocation error, $\Sigma_i \pi_i \varepsilon(\delta, f(\theta_i))$, is minimised numerically with respect to the cut-off point, k, given the prior probabilities, π_1 and π_2. The expressions (6.1-1 to 6.1-4) and (6.1-16) were evaluated for the training dataset over the whole sample space range of discrete states. Their frequencies for individual states completely defines the estimated empirical distribution of criterion values for both non-parametric models. The distributions are then sorted in ascending order of criterion value. Finally, all possible criterion values are systematically checked for possible candidates as cut-off (critical) points. The criterion value that results in the minimum value of $\Sigma_i \pi_i \varepsilon(\delta, f(\theta_i))$ is taken as the critical cutoff point to be used in the second phase on the test dataset. This procedure was repeated for a range of *30* prior probabilities in the range from *0* to *1* in equal steps. The actual prior probability was also included as a separate value.

Such consideration of the behaviour of an indirect discriminant procedure with respect to the prior distribution renders it as a typical Bayesian procedure. In chapter 4 it was pointed out that generally the direct procedures could also be viewed as Bayesian procedures with the adoption of a uniform prior. By contrast this example shows how a Bayesian approach may also be adopted for the class of indirect procedures.

Four medical datasets were used by Lack (1987) to test the procedures empirically, three relating to stillbirths and one to caesarean sections. The distributional distance method gave smallest overall error rates for all four datasets and largest values of average logarithmic score for two of the datasets. The centroid method gave largest overall error rates for three of four datasets. With increasing numbers of predictor variables the distributional distance method consistently resulted in smaller overall error rates than the other rules in all *4*

datasets. This pattern was stable over a wide range of prior probabilities. The work presented by Lack (1987) has the advantage that empirical datasets were employed and that the stability over a wide range of prior probabilities was taken as an extra performance measure. However, the datasets were to a large extent similar in nature and a generalisation to other types of dataset is limited in scope. Further, the absence of a comprehensive comparison with other established procedures such as the logistic model, nearest neighbour, kernel or interaction models also makes inferences difficult.

Improvements[11] on the specification of the generalised distance measure of Matusita (1955) and Goldstein and Dillon (1978) (expression 6.1-3) in terms of individual state specific weights related to absolute state frequencies, would seem helpful. This is because (6.1-12) depends only on relative frequencies and is therefore sensitive to variations in the p_{ij}.

Discrete multivariate data will generally lend themselves to presentation in the form of a state matrix with states $j=1,\ldots,s$ especially when the number of distinct levels per variable is low. This was the case in the previous example for the distributional distance procedure. In completing the subsection on distance models the work of Hills (1967) on deriving distance measures that are monotonically increasing functions of the number of discrete states s is of relevance in constructing discrimination procedures based on distance functions. Let Δ_s be a measure of discrimination between two multinomial distributions with s states and suppose the j^{th} cell of the multinomials are split into two cells j' and j'' to give multinomials with $(s+1)$ cells and a measure of discrimination Δ_{s+1}. Hills then stipulates four desirable properties that discrimination measures should possess:

---

[11] These are suggested in chapter 10 and lead to a modified version of Dillon and Goldstein's distributional distance procedure that is later used in comparative analyses.

(1) $\Delta_s \geq 0$ with equality if and only if the multinomial distributions are identical

(2) $\Delta_{s+1} \geq \Delta_s$

(3) $\Delta_{s+1} = \Delta_s$ if and only if the loglikelihood ratios in the cells j' and j'' are the same as that in the cell j

(4) Δ_s should be the sum of s contributions, one from each cell of the multinomial. It has the property that when the j^{th} cell is split into two cells j' and j'' to give multinomials with $(s+1)$ cells then the contributions of all cells except the j^{th} should be the same to both Δ_s and Δ_{s+1}.

Hills (1967) suggests

$$\Delta_s^{(2)} = \sum_{j=1}^{s} \frac{(p_{1j}-p_{2j})^2}{p_{1j}+p_{2j}} \; ; \quad (p_{1j}+p_{2j}\neq 1) \qquad (6.1\text{-}17)$$

as a measure satisfying all above conditions for two populations and gives a further expression for $g=3$ populations:

$$\Delta_s^{(3)} = \sum_{j=1}^{s} \left\{ \sum_{i=1}^{3} \frac{(p_{ij} - \frac{1}{2} p_j)^2}{p_j} \right\} \; ; \quad p_j=\textstyle\sum_i p_{ij} \qquad (6.1\text{-}18)$$

The use of 6.1-17 is demonstrated using a single dataset.

6.2 Recursive partitioning based procedures

Modern approaches are becoming more computer oriented. The increasing availability of computing power allows wider use of iterative *recursive partitioning* techniques. Use of classification trees in regression dates back to the *AID* (Automatic Interaction Detection) program developed at the Institute for Social Research, University of Michigan, by Sonquist and Morgan (1964). A dedicated classification program based on trees called *THAID* was later developed at the Institute for Social Research, University of Michigan,

by Morgan and Messenger (1973). Recently they have been replaced by the newer variants *classification and regression trees (CART)* proposed by Breiman, Olshen, Friedman and Stone (1984) and the *fast algorithm for classification trees (FACT)* due to Loh and Vanichsetakul (1988). Other variants of recursive partitioning procedures used for discriminant analysis applications include the *CHAID* algorithm described by Kass (1980) and the *ID3* and *C4.5* algorithms due to Quinlan (1993). Apart from that there exists also a range of software routines supplied by the major producers of statistical software packages such as the *TREEDISC*[12] set of macros of the *SAS* Institute (1996 β-Release, version 6.08).

Recursive partitioning algorithms iteratively search for optimal splits of the entire data for particular values of each variable. Optimality is defined in terms of maximising an heterogeneity criterion at each node of the classification tree. An example for the hypothetical data of figure 2.2-3 in chapter 2 is illustrated in figure (6.2-1)

---

12 *TREEDISC* is similar to the *CHAID* algorithm but differs from *CART*, which always forms two subsets, and from *ID3* or *C4.5*, which make every category a subset.

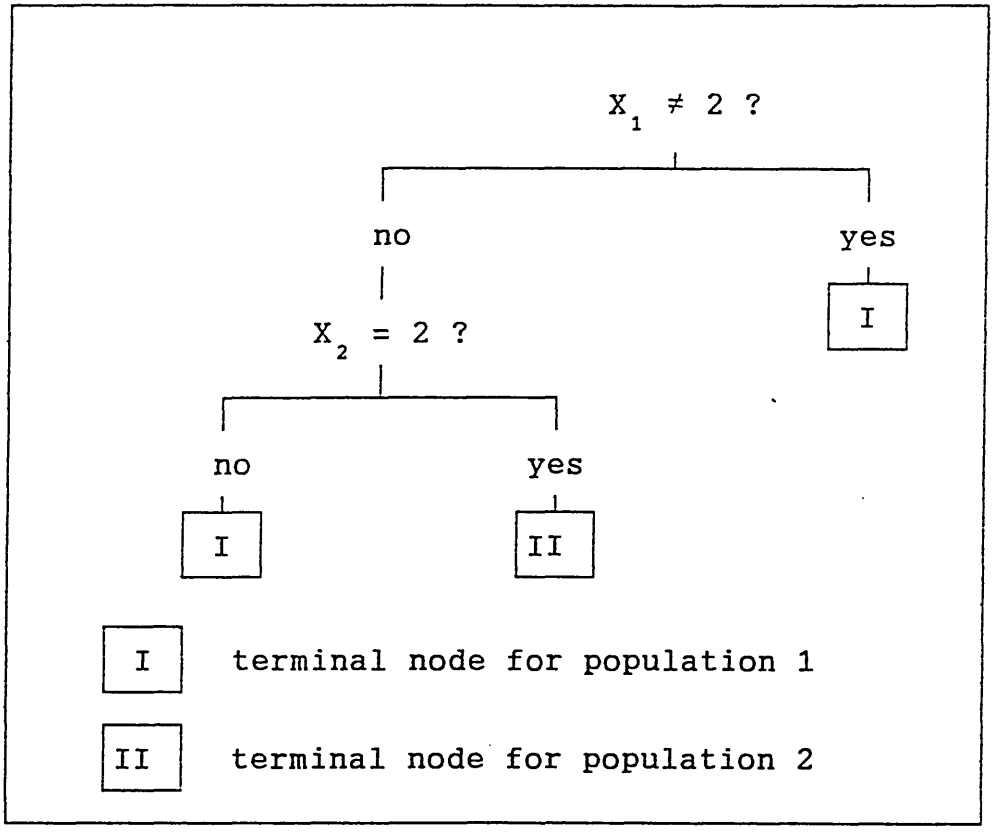


Figure 6.2-1: Decision tree with two-way splits

Separation is possible with recursive partitioning techniques for interactive data structures.

The example shows that it is possible to achieve complete separation of the training data by means of recursive partitioning providing the hierarchical trees are allowed to grow deep enough. A discriminant rule based on a completely grown tree using training data will be over specified when applied to test data. The problem particular to recursive partitioning is how to decide when to *stop* growing the tree.

A recently suggested procedure that is probably best described as a mixture of direct and indirect procedures is the *coupling method* suggested by Wernecke (1992). Providing several procedures are available and have been applied to a particular dataset the coupling method consists of simply averaging the allocation vectors obtained from all other

procedures thus giving the new vector. This pragmatic approach claims to produce allocations that are at least as good as any single procedure taken by itself.

6.3 Artificial neural network based procedures

The most recent developments in discriminant analysis are seen by the advent of so called *neural networks* or *artificial neural networks, ANN,* a technique that has been developed in the field of pattern recognition and machine learning since about 1985. Neural networks are commonly separated into *feed-forward networks,* also known as multi-layer perceptrons, used for classification and *symmetric recurrent networks,* known as attractor neural networks or Hopfield nets, used as associative memories. The first type is of relevance to discriminant analysis. General introductions to the theory of neural networks in classification may be found in Hertz et al (1991); Ritter, Martinez and Schulten (1991) and Ripley (1993, 1994). Weiss and Kulikowski (1991) provide a good elementary discussion of a variety of classification methods including statistical and neural methods. A good non-technical introduction to neural networks with a very balanced account of their usefulness in practical situations is given by Hinton (1992).

Artificial neural networks are used in three main ways:

(1) as models of biological nervous systems and "intelligence"

(2) as real-time adaptive signal processors or controllers implemented in hardware for applications such as robots

(3) as data analytic methods.

Here concern is with the latter. Artificial neural networks, like many statistical procedures, are capable of processing large amounts of data and are often reported to

produce good results. However this does not make them "intelligent" in the normal sense of the word. Artificial neural networks "learn" in much the same way as many statistical algorithms deal with estimation - but more slowly. If artificial neural networks are intelligent then many statistical methods must also be considered intelligent.

Many artificial neural network (*ANN*) models are similar or identical to popular statistical techniques such as generalised linear models, polynomial regression, non-parametric regression, discriminant analysis, projection pursuit regression, principal components and cluster analysis. This is particularly the case where the emphasis is on prediction of complicated phenomena rather than on explanation. The interest in artificial neural networks has boomed in recent years. Since 1990 accounts on neural networks in the context of discriminant analysis include those of Asoh and Otsu, 1990; Odom and Sharda, 1990; Webb and Lowe, 1990; Gallinari, Thiria, Badran and Fogelmansoulie, 1991; Reibnegger, Weiss, Wernerfelmayer, Judmaier and Wachter, 1991; Lowe and Webb, 1991; Garson, 1991; Kuhnel and Tavan, 1991; Brigatti, Filatov, Hoffman, Assad and Caprioli, 1993; Allen and Le Marshall, 1994; Curram and Mingers, 1994; Grozinger, Freisleben and Roschke, 1994; Kurita, Asch and Otsu, 1994; Osman and Fahmy, 1994; Mitra and Kuncheva, 1995; Wong, Jian and Taggart, 1995; and Sanchez and Sarabia, 1995.

Although many *ANN* models are similar or identical to well-known statistical models, the terminology in the *ANN* literature is quite different from that in statistics. Table 6.3-1 lists some equivalents where they exist. The statistical terms *sample* and *population* do not seem to have *ANN* equivalents. However the data are often divided into separate training and test sets.

neural network terminology	statistical terminology
features	variables
inputs	independent variables
outputs	predicted values
targets or *training values*	dependent variables
errors	residuals
training, learning adaptation or *self organisation*	estimation
patterns or *training pairs*	observations
(*synaptic*) *weights*	parameter estimates
higher order neurons	interactions
supervised learning or *heteroassociation*	discriminant analysis
unsupervised learning, encoding or *auto-association*	data reduction
competitive learning, or *adaptive vector quantization*	cluster analysis

Table 6.3-1: Terminology of neural networks

In the following, two examples of artificial neural networks for the discrimination problem are outlined using a graphical representation taken from Sarle (1994). In figures 6.3-1 and 6.3-2 circles represent observed variables and boxes represent values computed as a function of one or more arguments. The symbol inside the box indicates the type of *activation function*. Common activation functions are linear, logistic and threshold. Arrows indicate that the source of the arrow is an argument of the function computed at the destination of the arrow. Each arrow usually has a corresponding *weight* or parameter to be estimated. The parallel lines indicate that the

values at each end are to be fitted by least squares, maximum likelihood, or some other estimation criterion.

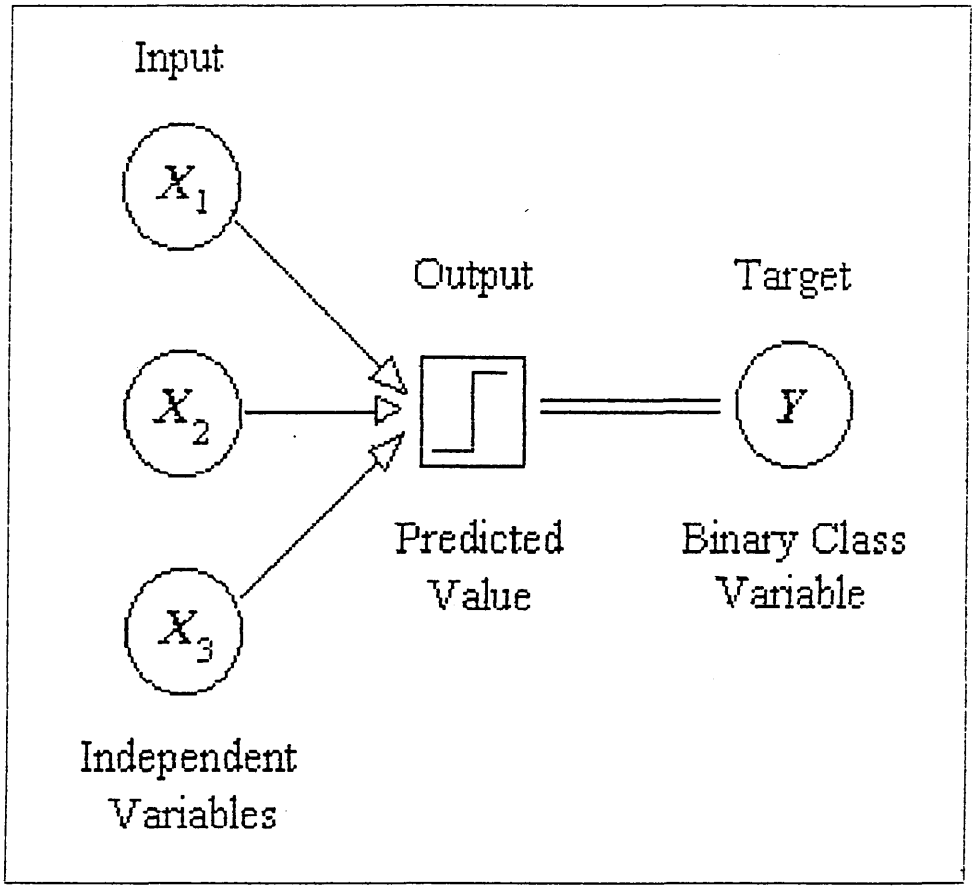


Figure 6.3-1: The "Adaline" perceptron

Figure 6.3-1 shows a simple neural network or perceptron (Rosenblatt, 1958) with a threshold activation function (represented by the step function in the box). This is equivalent to the linear discriminant function (Hand, 1981; Weiss and Kulikowski, 1991). A linear activation function would result in the multiple linear regression model and a sigmoid activation function leads to the multiple logistic regression model. The activation function in a perceptron is analogous to the inverse of the link function in a generalised linear model (*GLIM*) (McCullagh and Nelder, 1989).

In the graphical representation of the *Adaline* network in figure 6.3-1 the *threshold activation function*[13] maps a linear combination $\omega' x_j$ of independent variables X_{jk} onto a dichotomous response, $Y_j \in \{0,1\}$ depending on a cutoff c.

$$\hat{Y}_j = \begin{cases} 1 \text{ if } \omega' x_j > c \\ 0 \text{ otherwise} \end{cases} \qquad (6.3\text{-}1)$$

with $j=1,\ldots,n$ and $k=1,\ldots,q$. The vector of weights ω is chosen to minimise some fitting criterion, e.g. least squares

$$E = \sum_{j=1}^{n} (\hat{Y}_j - Y_j)^2 \ . \qquad (6.3\text{-}2)$$

Alternatives to least squares fitting of a function with target values $Y \in \{0,1\}$ have been suggested among others by Hinton (1992) and Ripley (1994). Figure 6.3-2 shows a two layer network with *logistic activation functions* in the middle *hidden* layer and a threshold activation function at the output. This multilayer perceptron is equivalent to a nonlinear discriminant analysis model. The classic algorithm for neural networks with at least one hidden layer is to take fixed steps in the direction of the steepest descent in minimising the criterion E,

$$\Delta\omega_{lm} = - \xi \frac{\partial E}{\partial \omega_{lm}} \qquad (6.3\text{-}3)$$

where ω_{lm} refers to the m^{th} weight in the l^{th} layer of a neural network. The derivatives of a fit criterion E with respect to the weights can be calculated recursively from output to input by using the chain rule, a technique known as *back propagation*.

---

[13] If instead of the threshold activation function (indicated by the step in the square box) a linear activation function is used a multiple linear regression model results and if a sigmoid or logistic activation function is used a nonlinear regression model is obtained.

Nonlinear regression models are represented as neural networks by introducing another so called *hidden layer* of activation functions. Their outputs are fed into other activation functions.

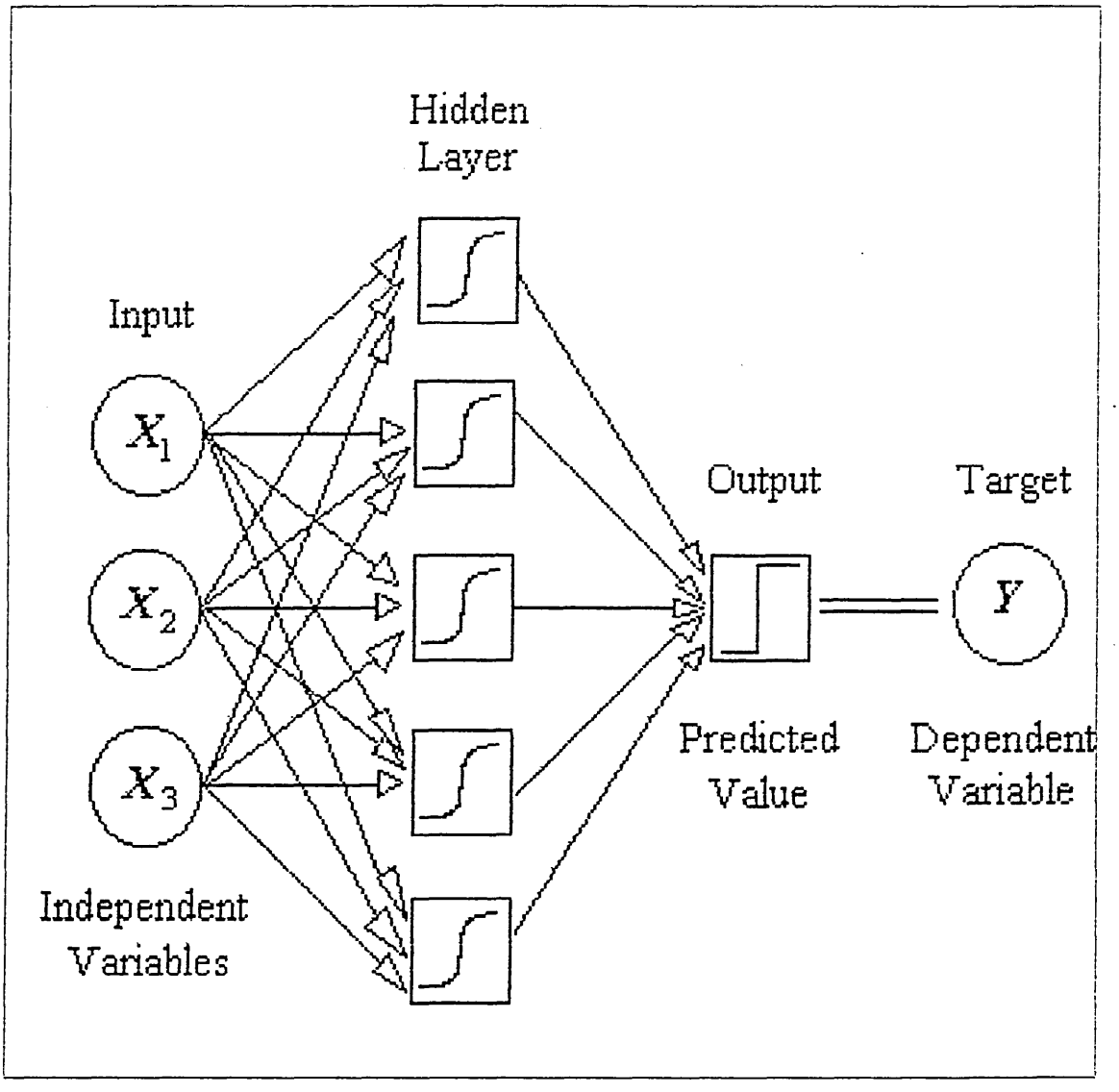


Figure 6.3-2: Multiple layer perceptron

Figure 6.3-2 shows such a multiple layer perceptron. The multiple input - single output network with a hidden layer of logistic response functions and a threshold output function is equivalent to the multiple nonlinear discriminant analysis model.

6.4 Use of *ANN's* as discrete discriminants

So far little work exists on actual applications of neural networks to problems of discriminant analysis dealing with discrete data. Examples where neural nets were applied generally include cases of continuous data and also where the classification problem had a high chance of being solved. Ripley (1994) lists some typical examples of such applications.

(1) grading of Danish bacon rashers

(2) botanical field guide key to species identification

(3) distinguishing male and female crabs of two species (Campbell and Mahon, 1974)

(4) recognising symbols on hand drawn maps (Hjort, 1986)

(5) predicting the occurrence of tsetse flies in Zimbabwe (Ripley, 1993)

A literature search conducted over the more recent years 1990 to 1995 yielded *25* reports of studies concerned with neural networks and discriminant analysis. However, of these studies only two deal with partially discrete data (Webb and Lowe, 1990; Yoon et al, 1993). In five cases of neural nets applied to continuous data (Asoh and Otsu, 1990; Gallinari et al, 1991; Garson, 1991; Grozinger et al, 1994; and Mitra and Kuncheva, 1995) superior performance is reported for the multilayer perceptrons equivalent to non-linear discriminant analysis. In the case of single layer perceptrons the discriminant only works if complete separation is possible between the populations.

What appears missing in the neural network literature concerned with discriminant analysis problems is a discussion of the validation issues with techniques such as crossvalidation or the bootstrap. In most reported cases input variables are continuous. Performance assessment of

ANN's applied to discriminant problems is at present still a comparatively new subject and, in the case of discrete data, virtually non-existent.

Figure 6.4-1 shows a simple landscape containing two local minima separated by a local maximum. Shaking can be used to allow the state *A* of the network (represented here by the ball bearing) to escape from a local minimum.

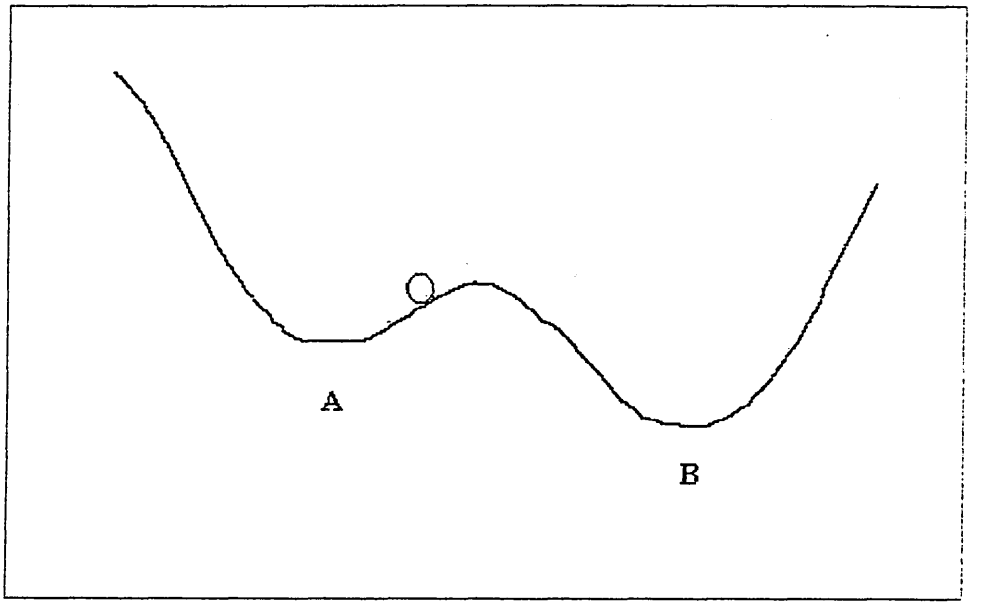


Figure 6.4-1: Local minima problems in *ANN's*

The expectation for the back propagation algorithm is that given a sufficient number of learning steps the overall error *E* may be reduced to an arbitrary small number. Two structurally inherent conditions, however, generally make that difficult. Firstly and fundamentally the mapping task must be solvable. With respect to the discriminant problems discussed so far, this is rarely the case - for all objects in the given samples and so for this reason alone *E* will not be able to be driven towards zero. Secondly because *E* is generally a highly complex function of all weights ω_{ab} this criterion will tend to exhibit several local minima. Depending on the particular choice of initial weights the steepest descent path will lead directly into an adjacent

minimum irrespective of how high this lies above the global minimum.

In spite of these difficulties the back propagation algorithm for multilayer networks is a considerable improvement on single layer networks, also known as perceptrons (Rosenblatt, 1958). Minsky and Papert (1969) have shown that provided a solution to the mapping or classification task exists the single layer perceptron will converge to that solution. The multilayer networks have the advantage that they can cope with more complex tasks.

Thus successful application of the *backpropagation* algorithm is generally only possible when the mapping problem is solvable. The typical examples of discriminant analysis for discrete data (see chapter 2), however, frequently show considerable overlap making an application of *ANN's* difficult. This suggests that *ANN's* may be particularly unsuitable for the type of dataset typical in discrete discriminant analysis.

One of the central issues in *ANN* research is the problem of dealing with local minima when seeking the global minimum. Figure 6.4-1 gives a physical representation of the problem in terms of a ball bearing placed in a landscape with several local minima.

The considerable number of weights or parameters requiring estimation in neural networks inherently bear the danger of overfitting the model. This, together with the other methodological difficulties mentioned above, indicates that applications of neural nets to problems of discrete discriminant analysis are generally fraught with difficulties.

6.5 Graphical techniques

It is an old principle in data analysis that where possible graphical plots ought to precede detailed analysis.

Frequently direct methods can reveal major structures of the data at an early stage. When there are more than *3* variables - and even with only *3* - the plots are technically difficult. Two popular methods for displaying differences between populations for multivariate data are Chernoff faces (1973) and harmonic curves (Ball and Hall, 1970; Andrews, 1972). Both methods reduce the multivariate data to stages that can be visualised in two dimensions. Inspection of these shapes may then aid discrimination. The techniques are also used for cluster analysis.

In its original form Chernoff allowed for up to *18* dimensions in a response vector. Each dimension became associated with one of *18* facial features. Bruckner (1978) has written a program to generate faces. Six facial features are incorporated in the construction: head, mouth, nose, eyes, eyebrows and ears. Shapes and sizes of these features are determined by values of the independent data. When data from distinctly different populations are used to generate such faces separation of the groups on the basis of visual inspection is readily done. Where differences become marginal, however, no formal procedure for distinguishing between faces exists and for this reason Chernoff faces are generally used solely for illustration rather than discriminative purposes.

In contrast the method of harmonic curves is more formal. For a given q-variate observation $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jq})$, $j=1,\ldots,n$, consider the corresponding curve

$$f(t;\mathbf{x}_j) = x_{j1}/\sqrt{2} + x_{j2}\sin t + x_{j3}\cos t + x_{j4}\sin 2t + x_{j5}\cos 2t + \ldots \qquad (6.5\text{-}1)$$

over the continuous interval $-\pi \leq t \leq \pi$.

In expression (6.5-1), deviating from standard notation, $f_{\mathbf{x}j}(t)$ stands for the function f of t for a given data point x_j and not for a statistical density as before. The formula above specifies n harmonic curves drawn in two

dimensions. Two data points are compared by visually studying the curves over $[-\pi,\pi]$.

Mardia, Kent and Bibby (1979) point out that the square of the L_2 distance

$$\int_{-\pi}^{+\pi} [f_x(t) - f_y(t)]^2 dt \qquad (6.5\text{-}2)$$

between two curves simplifies to

$$\pi \| \mathbf{x} - \mathbf{y} \|^2 \qquad (6.5\text{-}3)$$

which is proportional to the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ thus giving heuristic justification for the approach. In situations with large interpopulation differences resulting harmonic curves will be expected to form population specific patterns. As with the Chernoff faces the shapes of resulting curves are generally assessed visually. Further the resulting shape depends on the order in which variables are specified in (6.5-1), so care needs to be taken in construction of harmonic curves. Other ad hoc graphical techniques for illustrating multivariate data include so called star plots with q vertices, one for each of the q variables. The lengths of these vertices are drawn in proportion to values of the original data.

Comparatively little use is made of graphical techniques in applications of discriminant analysis for discrete data presumably because the discrete nature of the data leads to degenerate stages unsuitable for discrimination. From a mathematical point of view there is no reason for not using graphical techniques for discrete data.

Before concluding the review chapter mention should be made of a number of developments in computer graphics for representing multivariate data that are useful informal aids for classification. Schematic graphical displays of

multivariate observations proposed by Anderson (1957), Andrews (1972) and Chernoff (1973) can be and have been used for informal classification of objects. The essential idea is to represent either the individual training samples or some typical value (e.g. the mean of a group) via a schematic display, do the same for the test samples and then by inspection of these displays decide to assign a test sample to the group whose training sample displays (or typical value display) look "visually closest" to the display of the test case. In practice, large numbers of observations or variables, as well as poorly understood visual perception biases, can limit the usefulness of these graphical techniques.

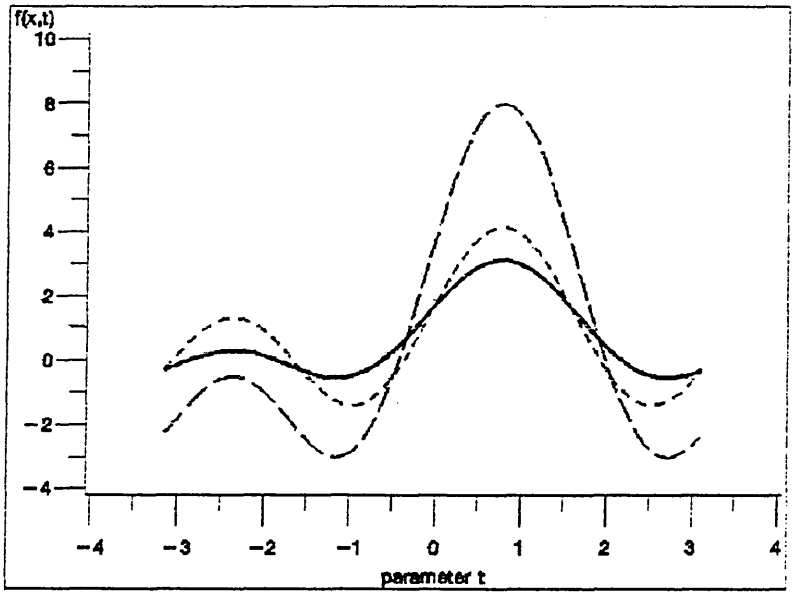


Figure 6.5-1: Harmonic curves

Figure 6.5-1 shows the representation of discrete multivariate data points $\mathbf{x}_j$ by harmonic curves (Ball and Hall, 1970; Andrews, 1972). Observations on each of the q variables are used as weights in a function $f(\mathbf{x}_j; t)$ as a linear combination of sines and cosines over the range $[-\pi \leq t \leq \pi]$. Three examples are shown for $q = 4$ with coordinates $(1,1,1,2)$, $(2,1,1,1)$ and $(1,3,3,3)$.

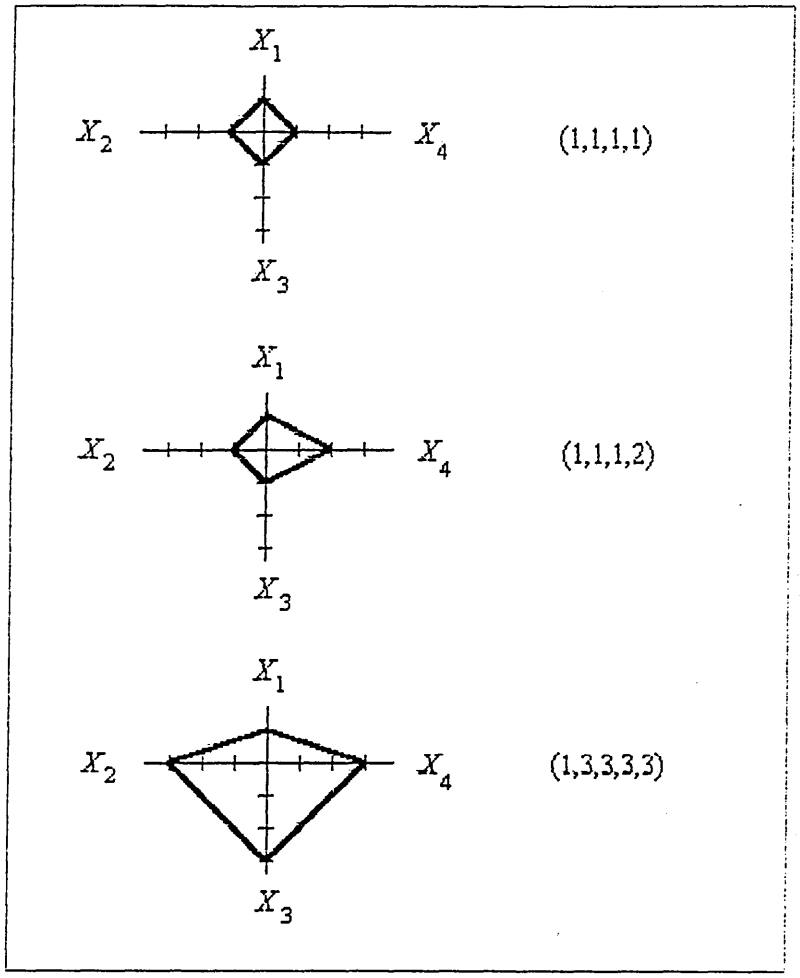


Figure 6.5-2: Star-plots

Figure 6.5-2 shows examples of star-plots for the three discrete *4*-variate observations displayed in figure 6.5-1. The star shapes are generated by mapping the observed values for each variable on one of the *q* - in this case *4* - axes radiating at equal angles from the origin. The end points are joined to produce the final shape. Star plots are mainly used for illustration rather than for formal discrimination problems.

6.6 Summary

Indirect discriminant procedures are more recent than direct procedures because of the generally computer intensive demands placed on execution. One may distinguish

distance based procedures, procedures based on *recursive partitioning*, *neural network* based procedures and *graphical* techniques.

Approaches to distance based procedures show that specification of the distance measures can be crucial. Thus this requires some knowledge of the data structure[14].

Graphical techniques are of little use for standard approaches to discriminant analysis mainly because of the large subjective element in graphical design. These techniques are probably better suited for methods of displaying multivariate data.

The neural network procedures, while currently increasingly popular mainly with protagonists of artificial neural networks, promise some applicability in discriminant analysis situations. However, research in *ANN*'s has so far concentrated on discriminant analysis applications for continuous data. Another difficulty with *ANN*'s is the problem of finding optimal solutions as it is frequently difficult to distinguish *local* from *global* minima in the *backpropagation* algorithm. Finally, the large number of parameters (*synapse strengths*) that have to be estimated bears the danger of overfitting of models.

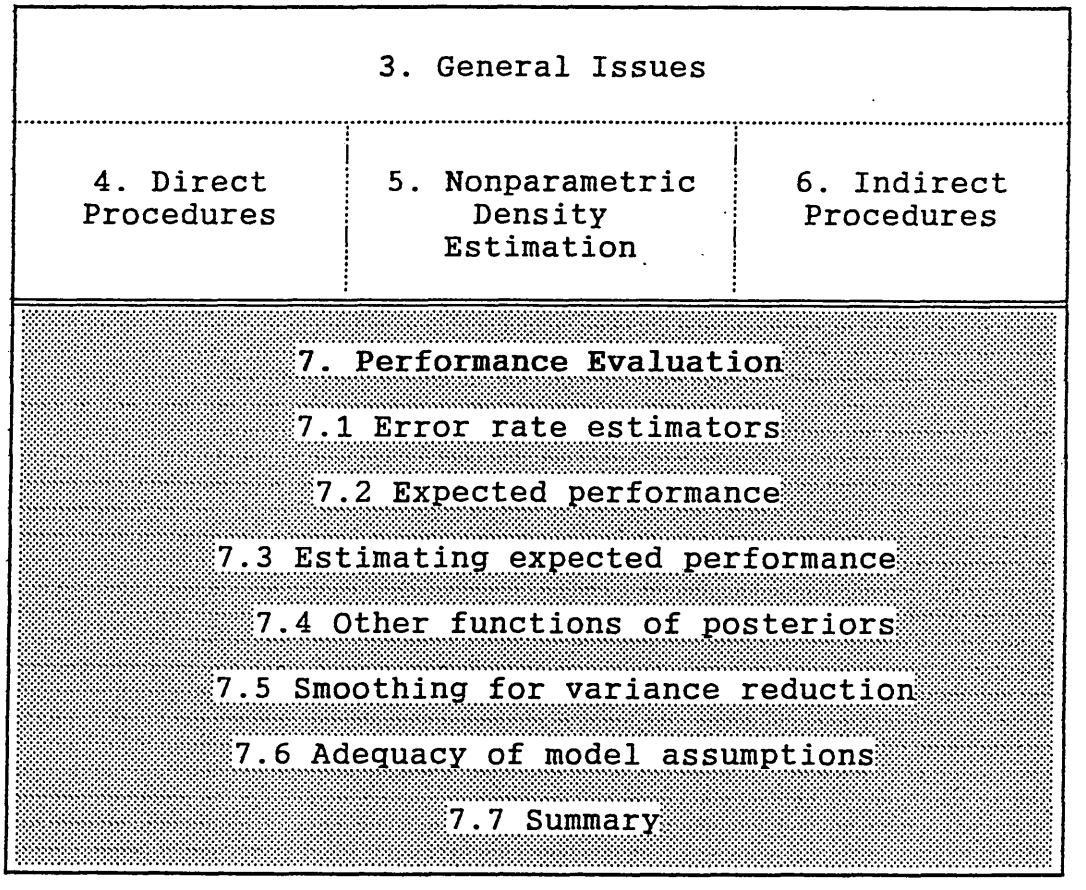
Major contributions on indirect discriminant procedures with relevance to discrete data include the work of Hills (1967), Minsky & Papert (1969), Goldstein & Dillon (1978), Breiman et al (1984), Hertz et al (1991), Ripley (1994).

---

[14] As seen in chapter 2 patterns in the data may lead to problems when using euclidean distances.

I: INTRODUCTION

II: REVIEW



III: METHOD

IV: RESULTS

V: DISCUSSION

The quality of a discriminant's performance is commonly expressed in terms of its *error rate* or *misallocation error*. Frequently a particular discriminant procedure is chosen because its error rate is less than that of other procedures[15]. Other desirable features of the error rate are low variance and low bias. Low variance means that repeated runs of a given discriminant on new samples will lead to stable error rate estimates while low bias implies that on average these estimates, more precisely their expected value, will be near to the true error rate for the given population.

Traditionally performance criteria for discriminant analysis are estimates of misallocation errors. They may be derived in one of three ways:

(1) "plug-in" methods: These substitute estimated parameters into theoretical expressions for the probability of misallocation.

(2) counting methods: The number of objects misallocated by a given discriminant procedure is divided by the total number of objects in the sample thus giving an error rate.

(3) posterior probability based methods: These less commonly used methods combine aspects of the above.

The last two categories exemplify additive estimation. The estimator can be viewed as an average of each sample observation's individual estimate of success probability.

The central problem in performance evaluation is that because of sampling errors uncertainty will generally prevail concerning the future performance of a rule derived

---

[15] Chapter 12 gives a detailed discussion of this and other factors influencing procedure selection.

from an initial set of data. Thus one has to make do with substitute estimators of the so called *optimum* or *true error rate*.

In addition to the common misallocation error other measures of performance such as error rate estimates based on posterior probabilities, performance scores derived from posterior probabilities and smoothing techniques for misallocation errors are occasionally reported (Glick, 1978; Titterington et al, 1981; Hora and Wilcox, 1982).

Further aspects of performance not centrally related to the present research concern the handling of costs of misallocation. It is common practice with direct procedures for discriminant analysis (see chapter 4) to adjust for this by shifting the cutoff point such that differential misallocation costs are proportionately accounted for. This adjustment adds a further term to the original discriminant function. In the case of the *2* group linear discriminant this term depends on a function of c_1 and c_2, the respective costs of misallocating an observation from populations Π_1 and Π_2. However, the basic operation of applying discriminant rules when adjusted for differential misallocation costs does not affect the general ideas pursued in this research. For this reason an investigation of misallocation costs is not pursued further.

In this chapter, section 7.1 outlines the different types of error rate (optimum, actual and apparent), section 7.2 introduces the notions of conditional and unconditional performance, section 7.3 discusses crossvalidation and bootstrap techniques, section 7.4 reviews the use of functions of posterior probabilities as performance estimators, section 7.5 looks at smoothing techniques for error rates, section 7.6 discusses the relevance of prior assumptions and model adequacy for performance estimation.

Optimality is generally defined in terms of the error rate. If for a given rule, δ_i, the misallocation error is smaller than or equal to that for all other rules, δ_j ($i \neq j$), then δ_i is considered optimal (Cochran and Hopkins, 1961; Glick, 1972). The inequality relating *theoretical optimum error*, *actual error* resulting from classification of future samples (Hills, 1966), and *apparent error* derived from classification of the learning sample, is central to estimation of bias and convergence of error rates. Hills showed that asymptotically the apparent error is always less than or equal to the theoretical optimum error and this in turn is always less than or equal to the actual error (see section 7.1 below).

Expressions for the *bias* of the actual and apparent error with corresponding correction terms (Cochran and Hopkins, 1961; Hills 1966; Goldstein and Wolf, 1977; Goldstein and Dillon, 1978) were developed later. Theorems required for proving the convergence of apparent error and actual error towards the theoretical optimum error with increasing sample size were provided by Glick (1972, 1973). Leaving-one-out or crossvalidation techniques (Lachenbruch, 1975) are important as they give virtually unbiased estimates of the optimum error given sufficient data. A crossvalidation option is available in the *SAS DISCRIM* procedure (1986).

The following error rates are usually considered:

(1) The *optimum error rate* - the rarely attained rate which would hold if the sample were unbiased and, in the case of parametric discrimination, when the distribution of the data including all parameters is known.

(2) The *actual error rate* - the rate which holds for a classification rule if it is used to classify future samples.

(3) The *apparent error rate* - the rate obtained by applying the discriminant rule to the training sample from which it was derived.

It is intuitive to expect the apparent error rate to be smaller than the actual error rate because the latter is derived from application of an estimated discriminant rule to a new and independent data sample. The apparent error rate benefits from the fact that the discriminant rule as well as the error rate are estimated from the identical data sample. The optimum error rate lies between apparent and actual error. This sequence is known to hold asymptotically and is generally expressed as a triple inequality:

$$\text{error}^{\text{apparent}} \leq \text{error}^{\text{optimum}} \leq \text{error}^{\text{actual}} \ . \qquad (7.1\text{-}1)$$

Many procedures that depend heavily on the assumption of normality have been proposed to estimate the error rates. Consideration is given here to estimators that may be used in any context. First, the apparent error rate simply classifies the training sample using the rule calculated from it. The estimator is typically optimistic and can badly mislead the user if the sample size is not much larger than the number of variables in the rule (Hills, 1966). It is also hazardous if there is initial misclassification in the training samples. However, for those cases in which the number of initially correctly classified observations is sufficiently large, the bias will be small.

In summary the apparent error is optimistically biased and should be used with caution when the sample sizes are small relative to the number of variables. The other methods mentioned can be useful alternatives in this case. Otherwise, the apparent error rate should be a satisfactory estimator. A first comprehensive bibliography on error rate estimators has been given by Toussaint (1974).

Two important approaches to performance evaluation that deal with obtaining useful criteria for evaluating discriminant performance must be mentioned as they are central to the present thesis. A detailed development of the basic ideas behind these approaches is however left to chapters 8 and 9, respectively.

The first is the concept of performance evaluation by exploiting information contained in the *posterior probabilities* of group membership. The *posterior error rate estimator*, initially proposed by Hora and Wilcox (1982), has the advantage of having lower variance and may be computed in conjunction with either of the aforementioned methods resubstitution, crossvalidation or bootstrap. In practical applications of discriminant analysis in other fields outside the theoretical statistical literature, however, the posterior error rate estimator is little used. For instance it was not quoted once in any one of *148* articles reporting results of discriminant analyses that were published in the medical literature in the years *1989*, *1991* and *1993* according to a *MEDLINE* literature research[16]. Instead it is evidently established practice to quote misallocation errors and to indicate methods of crossvalidation (see table 7.1-1). Use of separate training and test sets and leaving-one-out crossvalidation were the most commonly employed methods reported in the surveyed articles.

crossvalidation technique	*1989*	*1991*	*1993*	*total*
resubstitution	*30*	*46*	*48*	*124*
separate training and test sets	*6*	*3*	*4*	*13*
"leaving-one-out"	*6*	*3*	*2*	*11*
total	*42*	*52*	*54*	*148*

Table 7.1-1: Common crossvalidation methods

Table 7.1-1 details *148* research papers selected from the medical literature *MEDLINE* data base reporting studies

[16] Further details on this study are given in chapter 2.

involving discriminant analysis classified by year of publication and type of crossvalidation employed.

Secondly, the method of specifying a classification threshold as implemented in some statistical packages (e.g. *SAS*) has introduced a new way of looking at efficiency of discriminant procedures. The basic idea is that allocations are only made if the posterior probability exceeds a prespecified value. Similarly this intuitively appealing concept was also not used in any of the above mentioned articles searched from the *MEDLINE* database.

7.2 Expected performance

As one rarely has the situation where the group conditional distribution functions $F_i(x)$ are completely known one will have to resort to a sample t for estimation of both the discriminant rule and its error rate. Ideally one would like the rule and the error rate also to hold for future samples. Two situations need distinguishing: (1) application of the discriminant rule derived from t to future samples and (2) estimation of other discriminant rules from future samples. In the former case error rate estimates are averaged over future samples while in the latter case the discriminant rules themselves are averaged over future samples. The above distinction was originally drawn by Hills (1966) who coined the terms *conditional* and *unconditional probabilities of misallocation* to refer to the situations (1) and (2) respectively. In what follows the *actual* error rate frequently referred to in the literature will be used as equivalent to the *conditional* error. With the exception of McLachlan (1992) the distinction suggested by Hills (1966), although helpful in clarifying some of the often confusing terminology in error rate estimation, is rarely used in the literature on discriminant analysis.

The above concept will be referred to later in chapter 10 where the use of bootstrap methods for estimating the expected bias of misallocation errors is outlined.

The following notation is based on Hills (1966) and is also used by McLachlan (1992). The expressions $\varepsilon c_i(F_i,t)$ and $\varepsilon c(F,t)$ correspond to the actual error rates and $\varepsilon u_i(F)$ and $\varepsilon u(F)$ correspond to the expected error rates. Before the introduction of this terminology for the various types of error rates, there had been considerable confusion in the literature as pointed out by Cochran (1966). Let $\delta(\mathbf{x},t)$ denote an allocation rule formed from the training data t. Then the misallocation rates of $\delta(\mathbf{x};t)$, with respect to t, are defined by

$$\varepsilon c_{ik}(F_i;t) = \Pr\{\ \delta(\mathbf{x};t) = k \mid \mathbf{x} \in \Pi_i, t\ \} \qquad (7.2\text{-}1)$$

which is the probability, conditional on t, that a randomly chosen object from Π_i is allocated to Π_k $(i,k=1,\ldots,g)$. Then group specific conditional error rates are given by

$$\varepsilon c_i(F_i;t) = \sum_{i\neq k}^{g} \varepsilon c_{ik}(F_i;t) \qquad (7.2\text{-}2)$$

and the overall conditional error rate is given by

$$\varepsilon c(F;t) = \sum_{i=1}^{g} \pi_i\ \varepsilon c_i(F_i;t) \qquad (7.2\text{-}3)$$

The expected or unconditional error rates are obtained by taking expectations over the conditional error rates with respect to all training samples t:

$$\varepsilon u_{ik}(F) = E\{\varepsilon c_{ik}(F_i;T)\} = \Pr\{\delta(x,T) = k \mid X\varepsilon\Pi_i\} \qquad (7.2\text{-}4)$$

The group specific unconditional error rates are given by

$$\varepsilon u_i(F) = \sum_{i\neq k}^{g} \varepsilon u_{ij}(F) \qquad (7.2\text{-}5)$$

and the overall unconditional error rate by

$$\varepsilon u(F) = \sum_{i=1}^{g} \pi_i \, \varepsilon u_i(F) \qquad (7.2\text{-}6)$$

Methods for estimating conditional and unconditional expected performance are discussed in the following section.

7.3 Estimating expected performance

In section 7.1 it was reported that the apparent error based on resubstitution can exhibit considerable bias. The most effective way to reduce this bias would be to increase the sample size. This is not always practicable and the standard alternative approach is to employ some form of *crossvalidation* or data re-use. A popular choice is to split the data into separate training and test sets. This method has appeal because it is easy to implement and readily understood. In the following it will be referred to as the *hold-out* technique. However, especially with small data samples, the hold-out method can also exhibit considerable bias. Next the *leave-one-out* and the *leave-v-out* technique originally suggested by Lachenbruch (1975) are usually distinguished. By contrast they have the advantage of being almost unbiased.

The *leave-one-out* technique omits an observation, recalculates the classification rule from the remaining observations, classifies the deleted observation, and repeats these steps for each observation in turn. Counting the errors of misclassification yields an almost unbiased estimate of the error rate. Unfortunately, the variables indicating misclassification are correlated so that this estimate has comparatively large variance (McLachlan, 1992). In many cases, the mean square error of the leave-

one-out method is larger than that of the resubstitution estimator.

The *leave-v-out* technique is closely related to the leave-one-out technique. Initially, a suitable number k is found such that $k = n/v$ where k is an integer[17]. Thus the sample may be divided into k equally sized parts, each consisting of v observations. Then the first lot of v observations are removed from the sample and the discriminant rule is estimated from the remainder. Next the second lot of v observations is left out and so on until this estimation cycle has been repeated k times. On each cycle error rates are averaged. A popular choice of v is *10* percent of the sample size. When v equals *1* this reduces to the leave-one-out technique. Setting v to $n/2$ is equivalent to running the discriminant procedure twice with equal sized training and test sets. Provided enough data are available to carry it out, this also has the advantage of being nearly unbiased.

Crossvalidation techniques do not always provide low variance estimates of error rates (McLachlan, 1992). Two methods have been suggested to reduce the variance of error rate estimates. The first is the *posterior error rate estimator* (Hora and Wilcox, 1982) described in section 7.1 and later discussed in detail in chapter 8 and the second is a post hoc smoothing of the classical *counting error rate* (Glick, 1978) described in section 7.5. A third technique that can provide error rate estimators with lower variance is the *bootstrap* method (Efron, 1982). This seems to combine the best features of the previous two techniques: it is almost unbiased and it has a small variance. The major drawbacks of the bootstrap on the other hand are its expense and its inability, even asymptotically, to deal with sufficiently large biases in

---

[17] If this can't be achieved then either (a) n may be reduced at random to n' in order to enable integer division or (b) one of the left out lots of observations is adjusted from V to V' with a corresponding adjustment of the averaging process later on.

the original training sample (McLachlan, 1992). One must compute as many classification rules as there are replicates. If the classification rule is based on density estimation, this could become prohibitively expensive.

The *bootstrap* was introduced by Efron (1979), who investigated it further in a series of articles (Efron, 1982; Efron, 1983; Efron, 1987; Efron, 1990). The technique has become increasingly popular in statistics and is documented in the survey articles of Hinkley (1988), DiCiccio and Romano (1988) and Hall (1988) among others. The subject continues to be of interest as the very recent publications by Shao and Tu (1995) and Hall (1995) show. The bootstrap permits the estimation of the variability of a random quantity using just the given data.

An estimate $\hat{F}$ of the underlying distribution is formed from the observed sample. Conditional on the latter, the sampling distribution of the random quantity of interest is called the *bootstrap distribution* which gives an estimate of the true distribution F. The bootstrap can be implemented parametrically or nonparametrically (McLachlan, 1980; Schervish, 1981). In the former case the bootstrap data are generated with the vector of unknown parameters in the parametric form adopted for F_i ($i=1,\ldots,g$) replaced by an appropriate estimate formed from the original training data. In the latter case the bootstrap data are generated by using the empirical distribution function derived from the original data.

Computationally the nonparametric bootstrap is easier to handle, especially when dealing with comparative analyses including several discriminant procedures. By definition it is also ideal in the absence of distributional information. Simulation studies carried out by McLachlan (1980) show that there is little difference between mean square errors for both types of bootstrap indicating a high efficiency for the nonparametric bootstrap. Similar results were

obtained by Schervish (1981) in the case of $g \geq 3$ populations.

Figure 7.3-1 summarises the various crossvalidation techniques with respect to their approximate bias and variance properties. The diagram was constructed on the basis of the characteristics of various crossvalidation techniques as reported in the discriminant literature. Bias is plotted horizontally, approximate variance vertically. The variance of error rates derived from techniques based on partial or whole re-use of the data tends to be larger than that of the - albeit more expensive - bootstrap techniques. While the resubstitution technique shows negative, i.e. optimistic bias all others are positively, i.e. conservatively biased. The least bias results from application of leaving-one-out or the more expensive bootstrap techniques.

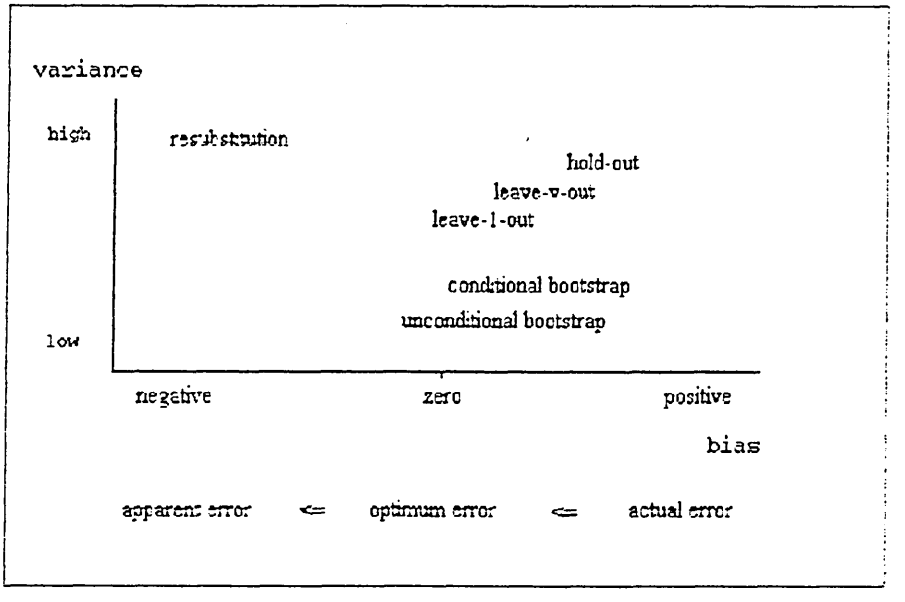


Figure 7.3-1: Characteristics of crossvalidation

7.4 Other functions of posteriors

Error rates are by far the most popular choice of *performance criteria* for discriminant analysis. The

misallocation error is usually computed by assessing differences between posterior probabilities *qualitatively*. In most applications an observation is allocated to the population with maximum posterior probability. The most commonly used performance criterion, the misallocation error, is very insensitive because it takes no account of the *relative seriousness* of different errors or of near misses - though it does have decision theoretic foundations such as in the field of the calculus of loss and risk functions (McLachlan, 1992). *Quantitative* assessments of the actual differences between posteriors by comparison are rare.

Other suggestions based on posterior probabilities are the *average logarithmic score, ALS,* and the *average quadratic score, AQS,* (Titterington et al, 1981) both of which are direct functions of the posterior probabilities.

$$ALS = -\ln\left(f(\Pi_i|X)\right) \qquad (7.4\text{-}1)$$

$$AQS = \left(1 - f(\Pi_i|X)\right)^2 + \sum_{i\neq k} f^2(\Pi_k|X) \qquad (7.4\text{-}2)$$

The *ALS* was used by Lack (1987) alongside customary error rates in a study of the behaviour of discriminants for varying values of prior probabilities. While the standard misallocation error tended to exhibit clearly identifiable minima for distinct intervals in the range of prior probabilities, the average logarithmic score showed a slightly more heterogeneous picture. This is seen as suggesting a capacity of performance criteria quantitatively based on posteriors for conveying additional information beyond that contained in the customary error rate.

7.5 Smoothing for variance reduction

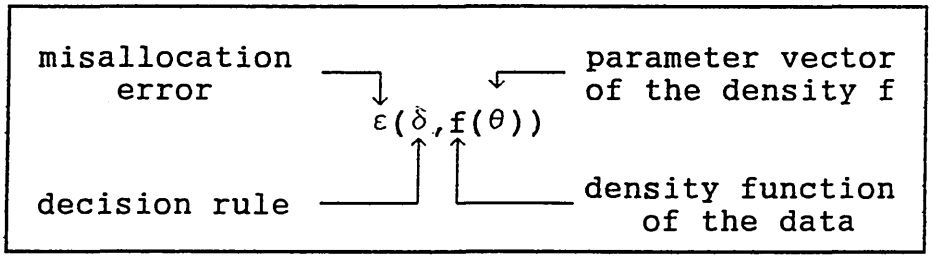
The second technique for reduction of variance of error rate estimators mentioned in section 7.4 consists of *post hoc smoothing* of the classical counting error rate introduced by Glick (1978). A feature of discrete data is that the number of combinations of variable values increases with the number of levels per variable. For q-variate dichotomous data the number of attainable discrete states is given by $s = 2^q$ and if all variables are trinomial $s = 3^q$ etc. Generally, if l_k is the number of levels of the k^{th} variable X_k, $s = \prod_{k=1}^{q} l_k$ where $k = 1,\ldots,q$.

For small values of l_k more objects will share identical feature vectors. This circumstance can be of considerable consequence for the misallocation error. To see this consider a vector of allocations $\delta(A,\mathbf{x})$ as a consequence of applying the discriminant procedure A to a given dataset $\mathbf{x}$. Next let $\delta(B,\mathbf{x})$ denote the corresponding allocation vector for procedure B. Let $\delta(A,\mathbf{x})$ and $\delta(B,\mathbf{x})$ differ only with respect to the j^{th} cell of the multinomial state vector. The resulting difference in error rates will be larger if the estimated population specific densities for this cell $\hat{f}(x_{ij})$ are substantial. This results in sudden sharp jumps in the error rate between procedures. To overcome this Glick (1978) suggested linear smoothing of the error rate.

7.6 Adequacy of model assumptions

The theoretical relevance of the correctness of underlying model assumptions is addressed by Victor (1976). However, comparatively little work on the consequences of model assumptions on performance has been reported. Relevant empirical evidence as well as basic theoretical background were provided by Christl and Stock (1973) and Victor (1976). They concluded that estimates of the actual error will converge to the optimum error only when the data model

underlying the discriminant procedure is true. For instance, if an independence model (see chapter 5 section 3) is fitted to discrete data that actually exhibit a first order interaction, an estimate of the actual misallocation error will not converge to the optimum error due to the inadequate data model assumptions. To make this distinction clear an appropriate indication of the actual model assumptions is suggested to clarify the information conveyed by standard error rates in the case of direct procedures. The following extends the work of Hills (1966) and Victor (1976). The usual error rate may be written as in figure 7.6-1 in terms of the allocation rule δ, the density function f and the parameter vector θ. The diagram shows components of the proposed extended definition for the error rate. When given with a hat, $\hat{}$, the arguments within the brackets indicate that the respective object is unknown and that either assumptions need to be made or that parameters require estimation.



(7.6-1)

Figure 7.6-1: Extended definition for errors

Generally only situations where either δ or θ have to be estimated are discussed. However, when uncertainty is allowed concerning the actual (assumed) distribution function $f(\mathbf{x})$ the relevance of model assumptions for the error rate becomes apparent. This is made clear in table 7.6-1 where the *3* classical types of error rate from section 7.1 are presented once under true model assumptions and once under false ones.

	data model assumptions underlying the discriminant procedure	
	true	false
optimum error	$\varepsilon(\delta, f(\theta))$	$\varepsilon(\delta, \hat{f}(\theta))$
actual error	$\varepsilon(\hat{\delta}, f(\hat{\theta}))$	$\varepsilon(\hat{\delta}, \hat{f}(\hat{\theta}))$
apparent error	$\varepsilon(\hat{\delta}, f(\theta))$	$\varepsilon(\hat{\delta}, \hat{f}(\theta))$

ε - error rate
δ - discriminant rule
f - assumed or true distribution of X
θ - parameter vector for X

Table 7.6-1: Errors using extended notation

The hat ^ symbol above the function symbol *f* in the second column indicates that errors are based on possibly inappropriate data models. In both cases - true or false model - the triple inequality "apparent error $\leq$ optimum error $\leq$ actual error" holds. The actual error is obtained when the discriminant rule derived from the training set is applied to new test data. The apparent error is obtained when the rule is applied to the training data. The optimum error requires complete distributional information. When the underlying data model is true both actual and apparent error can be shown to approach the optimum error asymptotically.

The expression in figure 7.6-1 distinguishes situations in which the data model is appropriate from situations where assumptions are made in the absence of prior information.

Christl and Stock (1973) showed that the use of appropriate data models in deriving discriminant procedures need not always lead to superior performance. Victor (1976) formalised these findings and suggested search strategies for selecting a discriminant procedure from within a *family*

of procedures of varying complexity. His findings are based on simulation studies of performance of discriminant procedures for a range of different sample sizes. Victor (1976) showed that for moderate sample sizes procedures based on alternative data models may be superior in terms of misallocation error. Figure 7.6-2 shows error rates under different data model assumptions. It is a modified version of a diagram originally given by Victor (1976) showing apparent, optimal and actual misallocation errors in relation to sample size under correct and false model assumptions, respectively. The original diagram by Victor (1976) displays *correct* classification rates as independent variables and involves several alternative model assumptions. For the sake of conformity with current notation and clarity of presentation his diagram had to be completely redrawn.

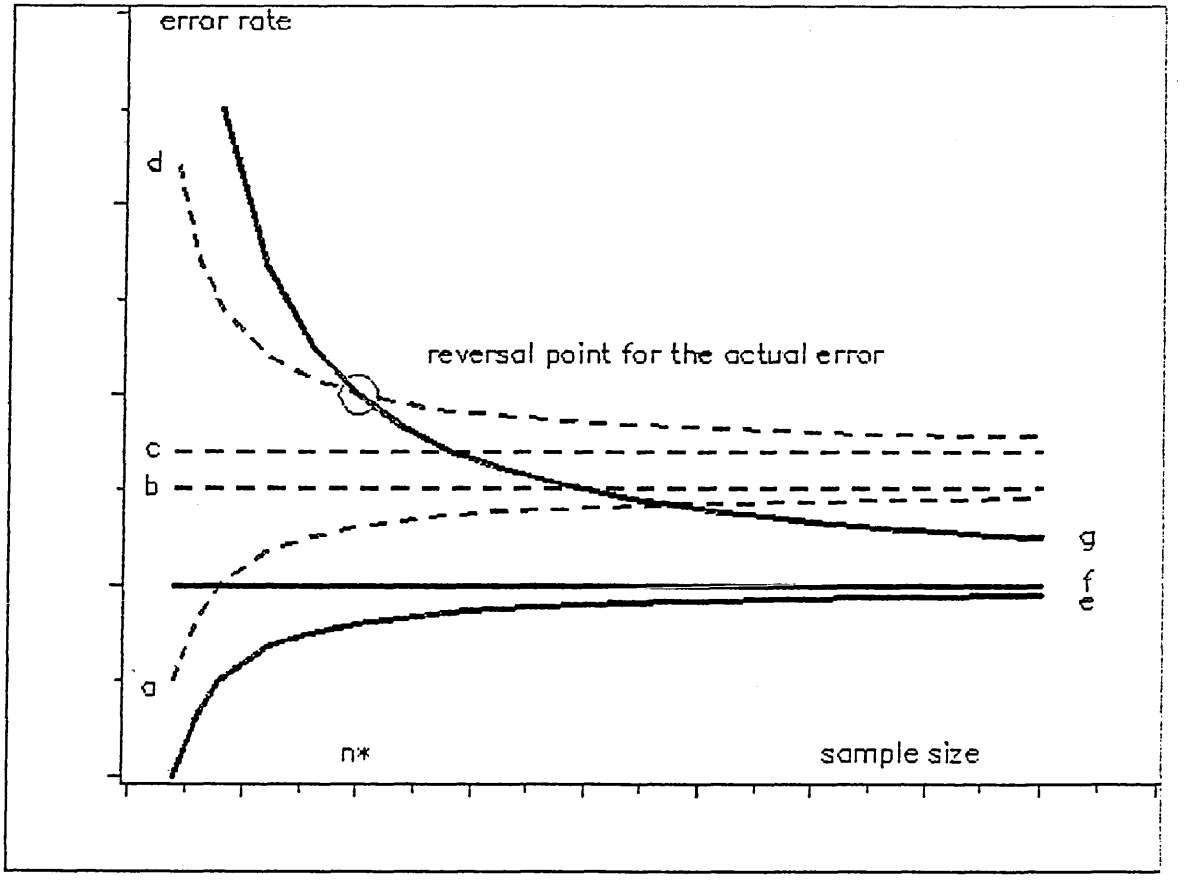


Figure 7.6-2: Error rates and model assumptions

The hypothetical diagram above shows apparent, optimum and actual misallocation errors as functions of sample size for different model assumptions. For the correct model estimates of the actual and apparent error (bold continuous lines labelled g and e) converge to the *same* optimum error (bold line f) with increasing sample size. For the alternative model the estimates (broken lines d and a) converge towards *different* asymptotes (broken lines c and b). With $n \longrightarrow \infty$ the apparent error for the incorrect model is minimised. When however the procedure based on the incorrect model is applied to new test data the actual error (broken line d) will not approach the optimum error $\varepsilon(\delta, f(\theta))$ (bold line f) but a false optimum (broken line c) and will show a bias, $\varepsilon(\delta, \hat{f}(\theta)) - \varepsilon(\delta, f(\theta))$, resulting directly from the assumption of an inappropriate model. Thus the differences (c - f) and (b - f) respectively correspond to the bias of the actual and apparent error rates.

In some situations discriminant procedures based on theoretically incorrect models may perform better than theoretically appropriate procedures. In the example this holds for sample sizes less than n^* because here the actual error based on the incorrect model assumptions (broken line d) is less than the actual error based on the correct model assumptions (bold line g).

The extended form of the misallocation error $\varepsilon\{\delta, f(\theta)\}$ unambiguously distinguishes situations in which the data model is appropriate from situations where assumptions are made in the absence of prior information. Figure 7.6-2 shows that for moderate sample sizes procedures based on alternative data models may be superior in terms of misallocation error.

7.7 Summary

The triple inequality stating that the *apparent error true* is at most equal to the *optimum error* which in turn is at most equal to the *actual error* constitutes the traditional basis of assessment of the bias of misallocation errors in discriminant analysis. The assessment of the variance of error rate estimators is of similar relevance. Simultaneous satisfaction of the joint requirement of low bias and low variance, and hence high stability, of error rates implies more elaborate estimation and crossvalidation techniques.

The concepts of *conditional* and *unconditional* performance are introduced as they are seen as central to crossvalidation techniques in general. As such they allow the reduction of bias and of variance. Other techniques used for variance reduction include the use of posterior probabilities and *post hoc smoothing* of the counting based error rate estimator.

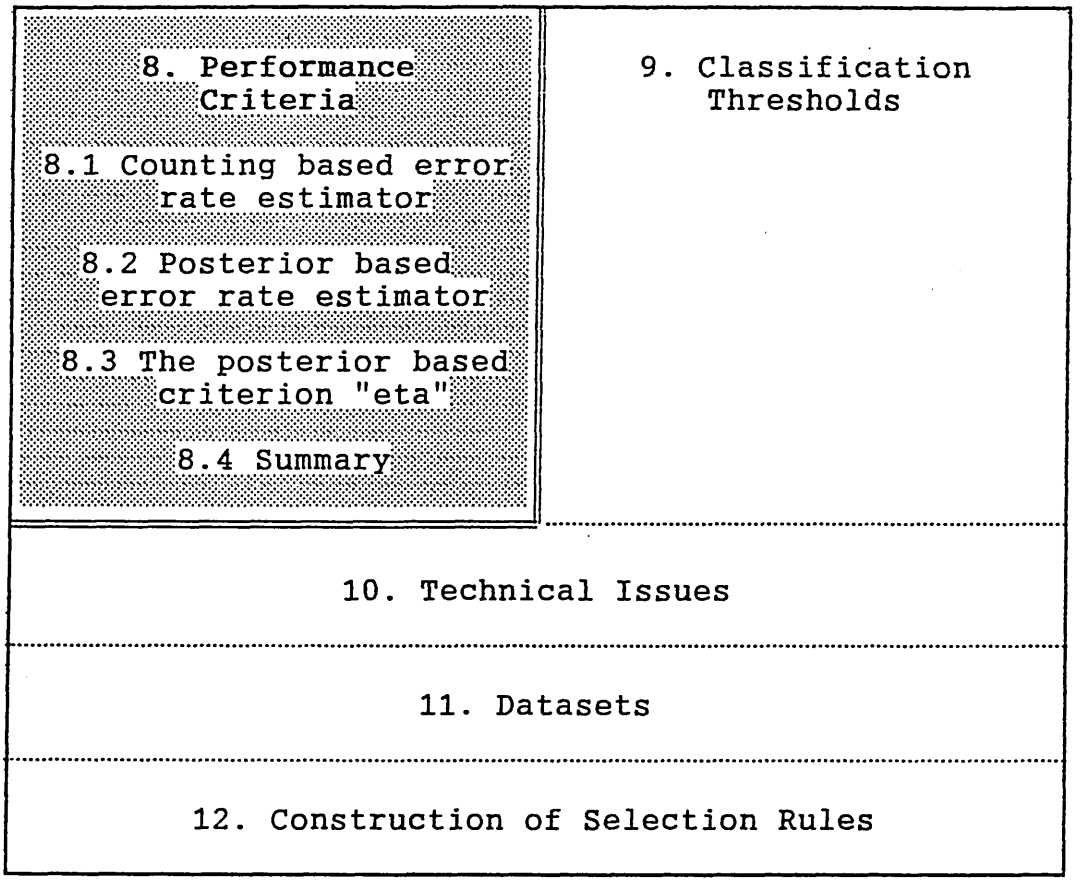
A sometimes underestimated factor of performance assessment concerns the adequacy of model assumptions. The quality of performance of a given procedure is shown to be additionally dependent on whether the distributional assumptions made are true or not. Under certain conditions a discriminant procedure based on incorrect assumptions may lead to better performance than a theoretically adequate procedure. A special notation is developed in order to distinguish performance under true model assumptions from performance under false model assumptions.

Major contributions on performance evaluation for discriminant procedures with relevance to discrete data situations include the work by Hills (1966), Victor (1976), Glick (1978), McLachlan (1980), Titterington (1981), Hora & Wilcox (1982), and McLachlan (1992).

I: INTRODUCTION

II: REVIEW

III: METHOD



IV: RESULTS

V: DISCUSSION

The three different performance criteria to be used for the comparative analyses of real and simulated datasets reported in chapter 13 are described in sections 8.1, 8.2 and 8.3. These are the common counting based misallocation error ($\varepsilon^{counting}$), the posterior based error rate estimator of Hora and Wilcox (1982) ($\varepsilon^{posterior}$) and the new eta criterion (η) respectively. For each of these suitable expressions that enable easy computation for the discrete data situation are derived. The expression derived for $\varepsilon^{counting}$ is given in terms of the *correct classification rate* in order to clarify its relationship with $\varepsilon^{posterior}$ which is also expressed in terms of the *correct classification rate*. The motivation for suggesting a new performance criterion (η) in section 8.3 is illustrated by examples of simple univariate discrete datasets.

8.1 Common "counting" based error rate

In the following an expression for the common counting based misallocation error $\varepsilon^{counting}$ is derived in terms of the *correct classification rate, ccr,* such that $\varepsilon^{counting} = 1 - ccr$. Assume the sample space Ω partitioned into g disjoint regions D_i, $i=1,..,g$, by the partition $\mathbb{D}=\{D_1, D_2, \ldots, D_g\}$ where D_i denotes the i^{th} region corresponding to the respective parent population Π_i. Further assume that *objects,* characterised by a multivariate independent q-variate discrete *feature vector* $\mathbf{X}$ are sampled from populations Π_i, $i=1,..,g$. Let $\mathbf{X}_i$ be distributed according to some multivariate discrete distribution specific to population Π_i such that $F_i(\mathbf{x}) = F(\mathbf{x}|\mathbf{x}\in\Pi_i)$ with corresponding population specific density function $f_i(\mathbf{x})$. Next assume that a discriminant rule $\delta(\mathbf{x})$ exists which partitions a given sample space $\{\mathbf{x}_j; j=1,..,n_i\}$ by $\mathbb{D}$ where $\sum_i n_i = n$ is the total number of objects in the whole sample $\{\mathbf{x}\}$. For the present assume that $\delta(\mathbf{x})$ is a discriminant rule such that

$$\delta\left[\ \mathbf{x}_j \,\middle|\, f(\Pi_i|\mathbf{x}_j)=\sup_k f(\Pi_k|\mathbf{x}_j)\ \right] = \Pi_i \qquad (\ 8.1\text{-}1)$$

with $k=1,..,g$. Thus $\delta(\mathbf{x})$ allocates a new object $\mathbf{x}_j$ to the population Π_i for which the estimated posterior probability of membership is greatest.

Let $\gamma_i(\mathbf{x})$ be an indicator variable such that

$$\gamma_i(\mathbf{x}_j) = \begin{cases} 1 \text{ if } \mathbf{x}_j \in \Pi_i \\ 0 \text{ otherwise} \end{cases}, \qquad (\ 8.1\text{-}2)$$

i.e. γ_i switches to *1* if the true population membership of $\mathbf{x}_j$ is Π_i. The *correct classification rate* of the discriminant rule $\delta(\mathbf{x})$ averaged over g populations is

$$\text{ccr} = \sum_{i=1}^{g} \pi_i \int \gamma_i(\mathbf{x})\ f_i(\mathbf{x})\ d\mathbf{x} \qquad (\ 8.1\text{-}3)$$

where π_i, $i=1,..,g$ are prior probabilities of population membership providing appropriate weights. For the present the integral notation indicates a continuous differentiable density $f_i(\mathbf{x})$ which is later replaced by the summation function for the case of discrete data. The customary - counting based - error rate estimator may next be expressed in terms of the *correct classification rate*:

$$\varepsilon_{\text{counting}} = 1 - \text{ccr} = 1 - \sum_{i=1}^{g} \pi_i \int \gamma_i(\mathbf{x})\ f_i(\mathbf{x})\ d\mathbf{x} \qquad (\ 8.1\text{-}4)$$

In the discrete data situation expression (8.1-4) may be calculated using

$$\varepsilon_{\text{counting}} = 1 - \sum_{i=1}^{g} \pi_i \sum_{j=1}^{s} \gamma_i(\mathbf{x}_j)\ \Pr\left\{\mathbf{X}_j=\mathbf{x}_j \,|\, \Pi_i\right\} . \qquad (\ 8.1\text{-}5)$$

where s is the number of discrete states of the data sample. As $Pr\{X_j=x_j|\Pi_i\} = n_{ij}/n_i$ (8.1-5) may also be written as

$$\varepsilon_{counting} = 1 - \sum_{i=1}^{g} \pi_i \frac{1}{n_i} \sum_{j=1}^{s} n_{ij}\, \gamma_i(x_j) \qquad (8.1\text{-}6)$$

where $\frac{1}{n_i} \sum_{j=1}^{n_i} \gamma_i(x_j)$ is the proportion of correctly allocated objects from population Π_i. Expression 8.1-6 may be used for purposes of practical evaluation of $\varepsilon_{counting}$ as defined by expression (8.1-4).

8.2 The posterior based error rate

In the following a corresponding expression is given for the posterior based error rate estimator of Hora and Wilcox (1982) again in terms of the *correct classification rate*. Let the prior probability π_i in the expression (8.1-3) of the preceding section be moved to form the term $\pi_i f_i(\mathbf{x})$ inside the integration sign $\int$. Then by Bayes theorem $\pi_i f_i(\mathbf{x})$ can be rewritten in terms of the posterior $f(\Pi_i|\mathbf{x})$ and the unconditional density $f(\mathbf{x})$ as

$$\pi_i f_i(\mathbf{x}) = f(\Pi_i|\mathbf{x}) \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}) = f(\Pi_i|\mathbf{x})\, f(\mathbf{x}). \qquad (8.2\text{-}1)$$

The summation sign in expression (8.1-3) can also be moved inside the integration sign and the correct classification rate (8.1-3) now becomes

$$ccr = \int \sum_{i=1}^{g} \gamma_i(\mathbf{x})\, f(\Pi_i|\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} \,. \qquad (8.2\text{-}2)$$

From this the *posterior probability based error rate estimator*, $\varepsilon^{posterior} = 1 - ccr$, is given by

$$\varepsilon^{posterior} = 1 - \int \sum_{i=1}^{g} \gamma_i(\mathbf{x})\ f(\Pi_i|\mathbf{x})\ f(\mathbf{x})\ d\mathbf{x} \qquad (8.2\text{-}3)$$

with the indicator function $\gamma_i(\mathbf{x})$ defined as before by (8.1-2), or equivalently

$$\varepsilon^{posterior} = 1 - \int \sum_{i=1}^{g} \max_k \left\{ f(\Pi_k|\mathbf{x}) \right\} f(\mathbf{x})\ d\mathbf{x}\ . \qquad (8.2\text{-}4)$$

Thus, (8.2-3) is calculated from the average of the maximum posterior probabilities for each observation. A feature of this estimate is that the prior probabilities of group membership do not appear explicitly - although they are introduced indirectly through the posterior probabilities $f(\Pi_i|\mathbf{x})$. Another feature of (8.2-3) is that its computation does not require knowledge of group membership which is evident from expression (8.2-4) where the indicator variable $\gamma_i(\mathbf{x}_j)$ has vanished and the joint density $f(\mathbf{x})$ has replaced $f_i(\mathbf{x})$. In the discrete data situation (8.2-3) may be computed from

$$\varepsilon^{posterior} = 1 - \sum_{i=1}^{g} \frac{1}{n_i} \sum_{j=1}^{s} n_{ij}\ \max_k \left\{ f(\Pi_k|\mathbf{x}) \right\} \qquad (8.2\text{-}5)$$

following a similar argument as for $\varepsilon^{counting}$ in expression (8.1-6).

The posterior probability based error rate estimator $\varepsilon^{posterior}$ tends to zero when $ccr \longrightarrow 1$ and vice versa. Desirable properties of $\varepsilon^{posterior}$ are its low variance characteristics (Fukunaga and Kessel, 1973; Hora and Wilcox, 1982; McLachlan, 1992). From equation (8.2-3) it can be seen that $\varepsilon^{posterior}$ depends only on the posteriors

for correctly allocated objects because $\gamma_i(x_j) = 1$ only if $x_j \in \Pi_i$.

8.3 The posterior based criterion η

In the following a criterion is suggested as an additional and possibly alternative performance measure for discriminant procedures. The new criterion, η, balances posterior probabilities for correct allocations *as well as* for misallocations. This is achieved by averaging posteriors with respectively chosen opposite weights. It is expected that by considering the *entire set of posterior probabilities* under certain circumstances η will be able to convey more information than customary error rate estimators such as $\varepsilon^{counting}$ and $\varepsilon^{posterior}$. In practical applications of discriminant analysis complete information (type of distribution, values of its respective parameters and prior probabilities) about the distributions $F_i(X)$ is rarely available. The following remarks state that in such situations the proposed symmetric η-criterion may be at least as good as the unilateral[18] posterior error rate estimator $\varepsilon^{posterior}$. Although designed within the context of finding selection guides for *discrete* discriminant analysis η is technically also applicable in *continuous* data situations.

Remark 8-1: *Let h^+ be the posterior probability $f(\Pi_i | x \in \Pi_i)$ of a correctly allocated object. Let h^- be the corresponding posterior probability $f(\Pi_i | x \notin \Pi_i)$ of a misallocated object. Let g be the number of populations sampled from. Then the set of posterior probabilities for misallocated objects and correctly allocated objects are asymptotically similar with $E(h^+) = E(h^-) = g^{-1}$ as the misallocation error ε increases.*

---

[18] *unilateral* because it uses only posteriors for correctly allocated objects.

Remark 8-2: *The set of posterior probabilities for correctly allocated objects h^+ and for misallocated objects h^- become asymptotically more heterogeneous with* the expected difference $E(h^+) - E(h^-) \longrightarrow 0.5$ *as the misallocation error* $\varepsilon \longrightarrow 0$.

Remark 8-3: *When the misallocation error* ε *is near zero then* $E(h^+)$ *and* $E(h^-)$ *under certain distributional configurations are independent.*

Remark 8-1 states that in datasets with little separation between populations the posteriors will tend to be of similar size. In the case of $g = 2$ populations $E(h) = 0.5$. But this also implies that under these conditions knowledge of $f(h^+)$ will also include knowledge about $f(h^-)$. As an illustration consider the univariate discrete 2-population dataset *A* shown in figure 8.3-1.

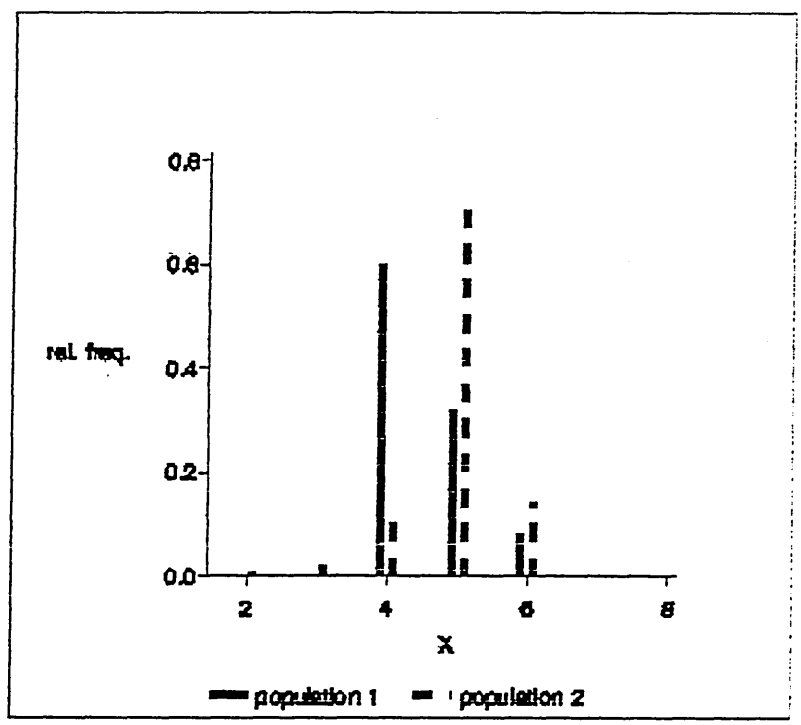


Figure 8.3-1: Conditional densities for data *A*

The data has been generated by *discretising* continuous distributions where $X \in \{2,3,\ldots,6\}$. The population specific distributions differ with respect to their means

and sign of skew. For the sake of distinguishing both populations in the histogram shown in figure 8.3-1 the horizontal positions of the columns for Π_1 have been shifted slightly to the left of respective discrete data points and those for Π_2 to the right. The actual data are given in table 8.3-1.

X	n_{1j}	n_{2j}
2	*0*	*1*
3	*0*	*2*
4	*60*	*12*
5	*32*	*71*
6	*8*	*14*
total	*100*	*100*

Table 8.3-1: Counts for dataset *A*

The linear discriminant function was applied to the dataset and the resulting sets of posterior probabilities extracted. These are shown in figure 8.3-2 separately for correctly allocated objects and for misallocated objects.

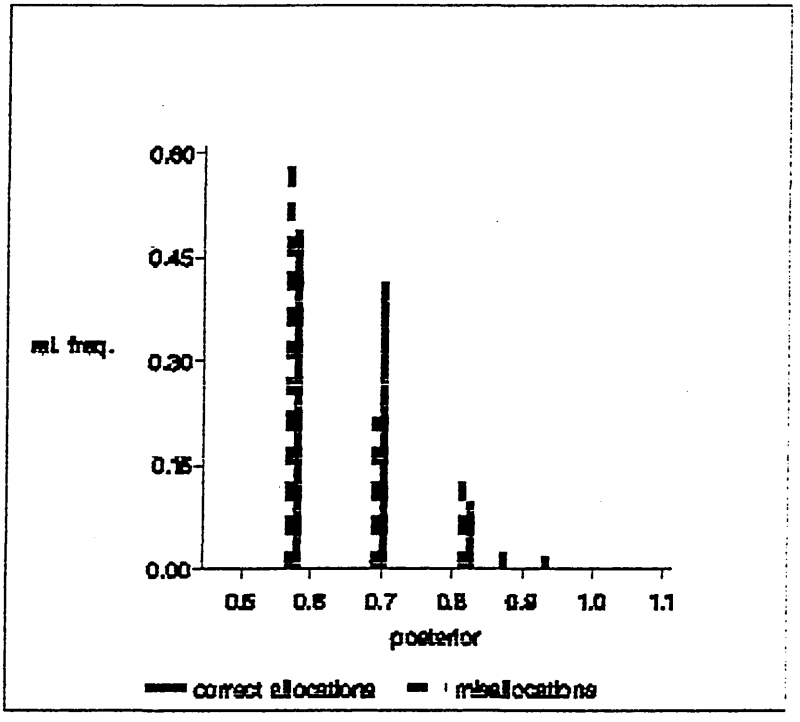


Figure 8.3-2: Posterior probabilities for data *A*

As to be expected from the large overlap between Π_1 and Π_2 in figure 8.3-1 the misallocation error $\varepsilon^{counting}$ is high (*0.275*). The cumulative posterior probabilities for $f(h^+)$[19] and $f(h^-)$[20] in figure 8.3-3 are correspondingly similar with the mean posterior for correctly allocated objects $E(h^+)$ = *0.649* and the mean posterior for misallocated objects $E(h^-)$ = *0.655*. These empirical findings support remark 8-1.

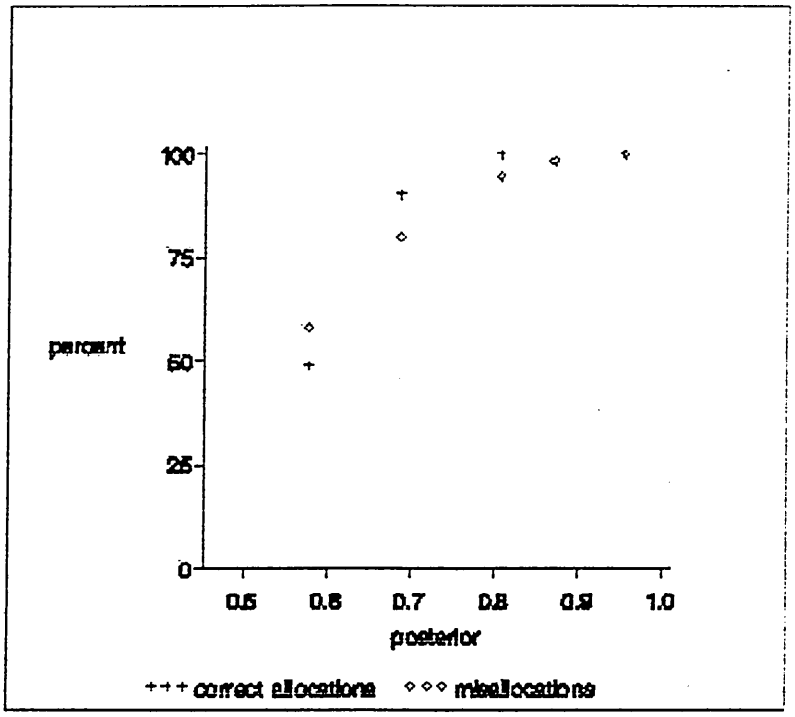


Figure 8.3-3: Cumulated posteriors for data *A*

Remark 8-2 states that as separability increases the values of h^+ and h^- will become more different with $E(h^+) > E(h^-)$ as greater doubt is cast on those (fewer) objects misallocated. Expressed differently, correct allocations are made with greater posteriors than misallocations; i.e. misallocations carry more doubt. As an illustration of this consider another univariate discrete example (data set *B*) shown in figure 8.3-4. The actual data are given in table 8.3-2.

---

[19] shown as crosses (+) in figure 8.3-3

[20] shown as diamonds (◇) in figure 8.3-3

x_j	n_{1j}	n_{2j}
2	*7*	*0*
3	*25*	*0*
4	*38*	*0*
5	*24*	*2*
6	*6*	*8*
7	*0*	*23*
8	*0*	*34*
9	*0*	*29*
10	*0*	*4*
total	*100*	*100*

Table 8.3-2: Counts for dataset *B*

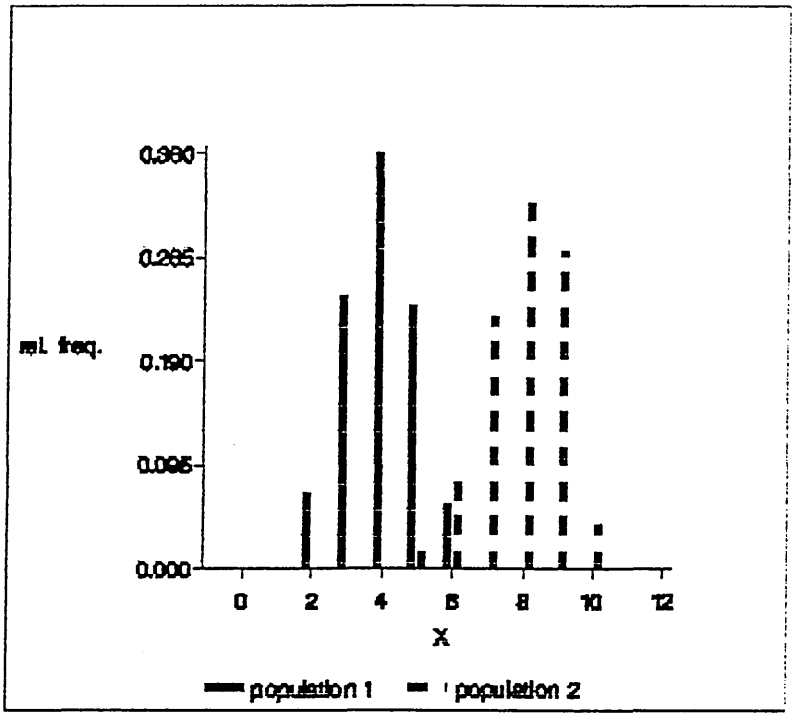


Figure 8.3-4: Conditional densities for data *B*

The distributions in figure 8.3-4 are again derived by discretising continuous densities with $X \in \{2,3,\ldots,10\}$. This time however the densities are symmetric with considerable difference in the population centroids. Consequently the misallocation error resulting from application of a linear discriminant function is low ($\varepsilon^{counting} = 0.04$). Correspondingly, the mean of the posteriors for correctly allocated objects ($E(h^+) = 0.974$) is considerably in excess of the mean posterior for

misallocated objects ($E(h^-) = 0.652$). This large difference in the distributions for h^+ and h^- is also evident from the cumulative posterior probabilities for $f(h^+)$[21] and $f(h^-)$[22] shown in figure 8.3-5. This supports remark 8-2.

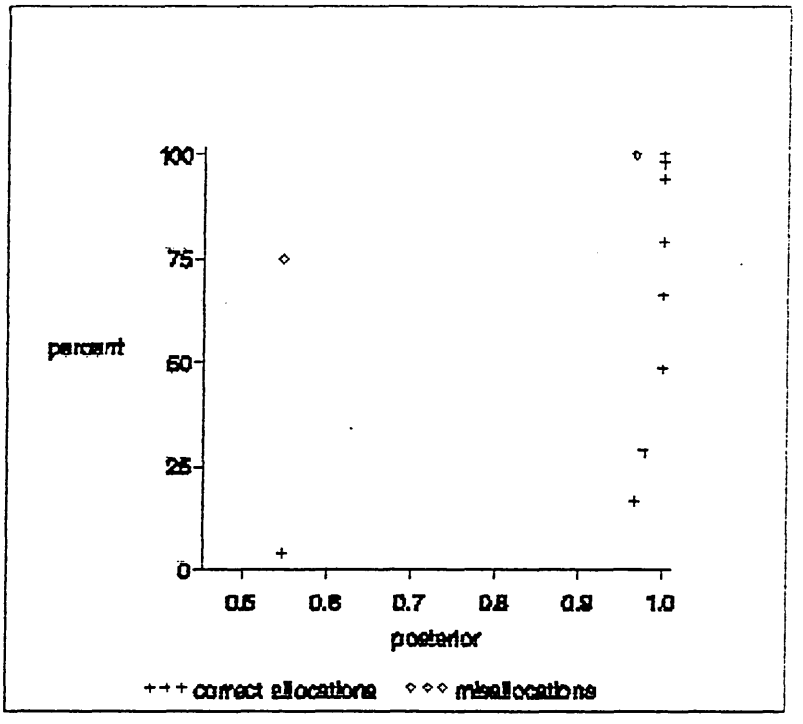


Figure 8.3-5: Cumulated posteriors for data *B*

From the example of dataset *B* it appears that to a certain extent knowledge of $E(h^+)$ implies knowledge of $E(h^-)$. This may be intuitively obvious as discriminant functions with high discriminatory abilities would be expected to exhibit correct allocations with large posterior probabilities yet misallocations with small posteriors near g^{-1}.

Remark 8-3, however, states that this is not the case. Under certain distributional conditions $E(h^-)$ is independent of $E(h^+)$. This is illustrated in the last example of dataset *C* (figure 8.3-6) where the misallocation error $\varepsilon_{counting} = 0.04$ is identical to that for the previous dataset *B*. The actual data are given in table 8.3-3.

---

[21] shown as crosses (+) in figure 8.3-3

[22] shown as diamonds ($\Diamond$) in figure 8.3-3

x_j	n_{1j}	n_{2j}
1	*0*	*1*
2	*0*	*0*
3	*0*	*0*
4	*71*	*1*
5	*19*	*0*
6	*2*	*1*
7	*3*	*0*
8	*3*	*1*
9	*1*	*5*
10	*1*	*1*
11	*0*	*11*
12	*0*	*79*
total	*100*	*100*

Table 8.3-3: Counts for dataset *C*

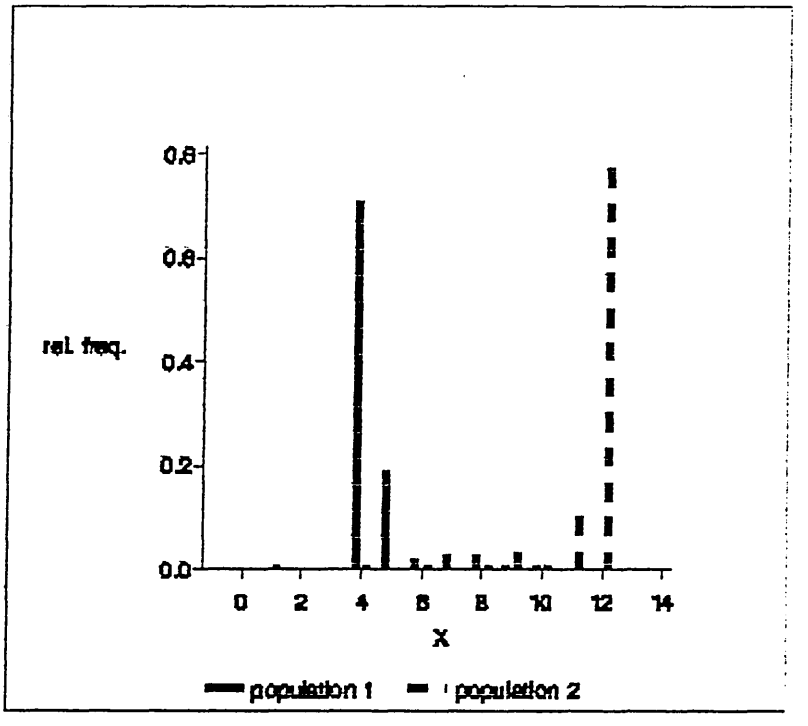


Figure 8.3-6: Conditional densities for data *C*

This time, however, the distributions are heavily skewed in opposite directions and the difference between means for populations Π_1 and Π_2 are twice as large as for dataset *B*. The mean of the posteriors for correctly allocated objects ($E(h^+) = 0.996$) is again of a similar order of magnitude as in example *B*. But instead the mean posterior for misallocated objects ($E(h^-) = 0.814$) is considerably larger

than in example *B* as can be seen from the plot of cumulative posterior probabilities in figure 8.3-7.

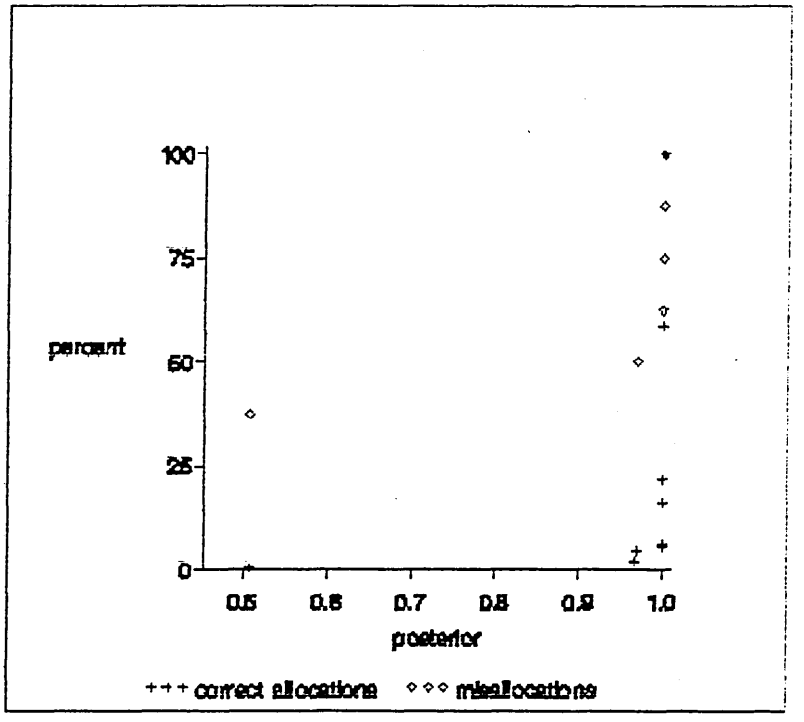


Figure 8.3-7: Cumulated posteriors for data *C*

The fact that while $\varepsilon^{counting}$ is identical in examples *B* and *C* and that $E(h^+)$ is also similar, yet $E(h^-)$ has increased by *25* percent demonstrates that under certain conditions a performance criterion based on the entire distribution of posterior probabilities $f(h)$[23] can contain more information than a criterion based only on $f(h^+)$. This supports remark 8-3.

Following on from remark 8-3 the criterion η may be constructed. This is achieved by extending definition (8.1-2) for the indicator variable $\gamma_i(\mathbf{x}_j)$ used in defining $\varepsilon^{counting}$ in section 8.1 and introducing the symmetric indicator variable $\xi_i(\mathbf{x}_j)$. Let

$$\xi_i(\mathbf{x}_j) = \begin{cases} +1 \text{ if } \mathbf{x}_j \in \Pi_i \\ -1 \text{ otherwise} \end{cases} . \qquad (\ 8.3\text{-}1)$$

[23] $f(h)$ is the distribution of posterior probabilities across S discrete states of the data sample.

In analogy with the integral in (8.2-2) and now using the extended definition for the indicator variable $\xi_i(\mathbf{x}_j)$ define for the i^{th} population the quantity

$$\eta_i' = \int \xi_i(\mathbf{x})\ f(\Pi_i|\mathbf{x})\ f_i(\mathbf{x})\ d\mathbf{x} \qquad (\ 8.3\text{-}2)$$

and next its corresponding weighted sum over all g populations

$$\eta' = \sum_{i=1}^{g} \pi_i\ \eta_i' \ . \qquad (\ 8.3\text{-}3)$$

Expression (8.3-3) gives the raw unadjusted eta criterion η'. This statistic takes on values in the range $\{-1,+1\}$. For the sake of comparing its bias and variance characteristics with those of $\varepsilon^{counting}$ and $\varepsilon^{posterior}$ the raw η' is further transformed (see chapter 10) to lie in a comparable range. As stated at the beginning of section 8.3 the η criterion is not to be viewed as a substitute for the error rate but rather as an additional more general criterion of performance. Hence in practical applications its main use is seen as augmenting other estimates of the misallocation error to enable reliability assessment of discriminant procedures applied to discrete data.

For practical purposes η' may be computed from

$$\eta' = \sum_{i=1}^{g} \pi_i \sum_{j=1}^{s} \xi_i(\mathbf{x}_j) f(\Pi_i|\mathbf{x}_j) Pr\{\mathbf{X}_j=\mathbf{x}_j|\Pi_i\} \qquad (\ 8.3\text{-}4)$$

which in analogy to corresponding expressions for $\varepsilon^{counting}$ and $\varepsilon^{posterior}$ is

$$\eta' = \sum_{i=1}^{g} \pi_i \frac{1}{n_i} \sum_{j=1}^{s} n_{ij} \xi_i(\mathbf{x}_j) f(\pi_i|\mathbf{x}_j) \qquad (\ 8.3\text{-}5)$$

Another way of looking at η is achieved by rewriting the components η_i' from expression (8.3-2) as

$$\eta_i' = \int_{R_i^+} f(\Pi_i|\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x} - \int_{R_i^-} f(\Pi_i|\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x} \qquad (8.3\text{-}6)$$

where R_i^+ and R_i^- indicate regions of correct and incorrect allocation with respect to the distribution of **X** in population Π_i. Let $E_{R_i^+}[.]$ and $E_{R_i^-}[.]$ indicate expectations in the regions R_i^+ and R_i^- with respect to the conditional densities $f_i(\mathbf{x}) = f(\mathbf{x}|\Pi_i)$. Then expression (8.3-6) is equivalent to

$$\eta_i' = E_{R_i^+}\left[f(\Pi_i|\mathbf{x}) \right] - E_{R_i^-}\left[f(\Pi_i|\mathbf{x}) \right] \qquad (8.3\text{-}7)$$

or

$$\eta_i' = \frac{n_i^+}{n_i} \bar{h}_i^+ - \frac{n_i^-}{n_i} \bar{h}_i^- \qquad (8.3\text{-}8)$$

where $\bar{h}_i^+$ and $\bar{h}_i^-$ are shorthand notations for the average posteriors within the respective regions R_i^+ and R_i^-. n_i^+ and n_i^- are respectively the number of correctly and incorrectly allocated objects. The components η_i' may thus be considered as *the proportion of correctly allocated objects times the average of corresponding posteriors minus the proportion of incorrectly allocated objects times the average of corresponding posteriors within population* Π_i. This definition clearly shows the conditions that will maximise η_i' and hence η': low proportion of misallocated objects and large *heterogeneity* between the distribution of posteriors in R_i^+ and R_i^-

8.4 Summary

The common misallocation error is computed by counting the number of objects misallocated by a discriminant rule and dividing this by the total number of objects in the sample. As this approach is based on counting the term $\varepsilon^{counting}$ is introduced. An expression in terms of an indicator variable, $\gamma_i(\mathbf{x}_j)$, that switches between *0* and *1* for incorrect and correct allocations, and the conditional densities, $f_i(\mathbf{x})$, is developed for $\varepsilon^{counting}$.

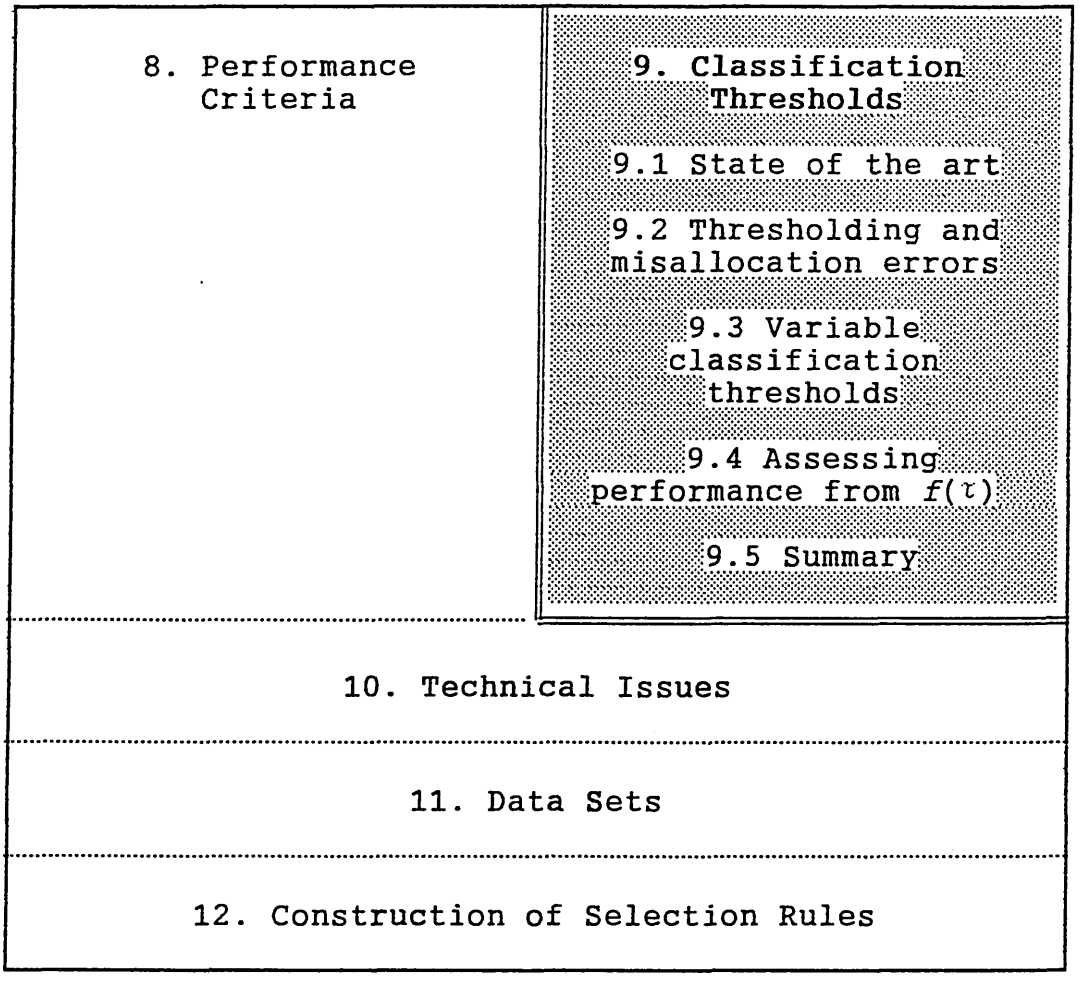
An analogous expression based on the same indicator variable, $\gamma_i(\mathbf{x}_j)$, is given for the posterior probability based error rate estimator of Hora and Wilcox (1982). This is called $\varepsilon^{posterior}$. As $\gamma_i(\mathbf{x}_j)$ switches to *0* for incorrect allocations, not all posteriors enter $\varepsilon^{posterior}$.

Three remarks are postulated stating that under certain conditions the information contained in the entire distribution of posterior probabilities is greater than the information contained in the subset distribution of posteriors for correctly allocated objects. The properties stated in these three remarks are illustrated using different hypothetical datasets. Based on this a new performance criterion, η, is constructed, however this time using an indicator variable, $\xi_i(\mathbf{x}_j)$, taking on the values *-1* and *+1* for misallocations and correct allocations, respectively. As a consequence the entire distribution of posteriors enters into the new performance criterion, η.

I: INTRODUCTION

II: REVIEW

III: METHOD



IV: RESULTS

V: DISCUSSION

The reasons for introduction of classification thresholds are twofold. The first is to increase the reliability of a discriminant rule and the second is to make selection from a given set of rules easier. These two points are briefly discussed in the following.

(a) Assume that the joint densities for two discrete states X_j and X_k are the same such that $f(x_j) = f(x_k)$. If the posterior probabilities computed for a given discriminant procedure differ considerably, e.g. if for the j^{th} state $f(\Pi_1 | x_j) = 0.10$ and $f(\Pi_2 | x_j) = 0.90$ then - assuming correct model assumptions - great confidence would be placed on allocation to population Π_2. On the other hand if differences between posteriors are only marginal such that, for instance, for the k^{th} state $f(\Pi_1 | x_k) = 0.49$ and $f(\Pi_2 | x_k) = 0.51$ then more doubt would be placed on allocation to population Π_2. This is so because a further test sample might lead with greater probability to posteriors for the k^{th} state that are reversed due to sampling variability. Given that the joint density $f(x)$ is the same for both states j and k sampling is more likely to affect the sign of the smaller difference between posteriors. One way to avoid such misallocations is to specify a minimum *classification threshold*, τ_{min}, that must be exceeded by a posterior $f(\Pi_i | x)$ in order to qualify for an allocation. The larger the threshold the greater the degree of confidence in the allocation and thus the greater the reliability of the discriminant rule. The price for greater reliability, however, is the consequently larger proportion of rejected (not classified) cases.

(b) Assume that for a given discrete dataset with $g = 2$ populations two different direct discriminant procedures have been applied and that for each discrete state X_j corresponding sets of posteriors are available. Further assume that the allocation rules and thus the misallocation errors for both procedures are identical. A sufficient condition for this under direct discriminant procedures[24] is that the *signs of differences* between posteriors are identical for both procedures. The more *reliable* procedure in the above sense (a), however, should also exhibit larger posterior differences. This will show up when classification thresholds are introduced such that the misallocation error rises more slowly for the more reliable discriminant procedure. Conversely, less reliable discriminant procedures will exhibit faster rises in misallocation errors with increasing classification threshold, τ. This feature will be developed to form a further basis for procedure selection in addition to inspection of the performance criteria discussed in chapter 8.

The present chapter deals with the consequences of using classification thresholds. The current state of the art is reviewed in section 9.1. Section 9.2 illustrates the relationship between classification thresholds and the misallocation error by means of two real discrete datasets. In section 9.3 the concept of fixed classification thresholds is extended to *variable classification thresholds*. Section 9.4 suggests various ways of using the empirical distribution of relative differences between posteriors in evaluating performance of discriminant procedures.

---

[24] i.e. when allocation is made to the population with the larger posterior probability computed from group conditional densities and given prior probabilities.

Application of direct discriminant procedures stipulates that allocations are made when the likelihood ratio differs from unity *irrespective* of the actual size of the difference from *1*. Yet it appears natural that one's confidence in having made a correct allocation would be related to the *absolute size* of the maximum posterior.

When posterior probabilities only barely exceed the critical sizes required for allocation - in the case of $g = 2$ with $\pi_1 = \pi_2$ this is *0.5* - doubt may be cast on the reliability of such *marginal* allocations. To control this one may use *allocation thresholds*. Let $h_i = f(\Pi_i|\mathbf{x})$ and $h^{(1)} = \max_i\{f(\Pi_i|\mathbf{x})\}$ be shorthand notations for the posterior probability of Π_i given that an observation X takes on the value x and the maximum posterior over g populations ($i=1,\ldots,g$) respectively. Consider as an illustration the hypothetical univariate data example presented in table 9.1-1 for two populations with discrete $X \in \{1,2,\ldots,10\}$ and slight opposite skew.

X	f_1	f_2	$h^{(1)}$
1	*0.19*	*0.00*	*1.00*
2	*0.38*	*0.00*	*1.00*
3	*0.25*	*0.00*	*1.00*
4	*0.10*	*0.01*	*0.91*
5	*0.04*	*0.03*	*0.57*
6	*0.02*	*0.03*	*0.60*
7	*0.01*	*0.07*	*0.88*
8	*0.01*	*0.26*	*0.96*
9	*0.00*	*0.42*	*1.00*
10	*0.00*	*0.19*	*1.00*

Table 9.1-1: Classification threshold example

The second and third columns of table 9.1-1 give the conditional densities for Π_1 and Π_2. The fourth column gives the maximum posterior $h^{(1)}$ derived from application of a discriminant rule based on the multinomial model (chapter 4, section 1). In terms of the state matrix

notation the posterior probabilities then take the simple shape $h_{ij} = p_{ij} / (p_{1j}+p_{2j})$, $i=1,2$, where p_{ij} are the relative cell frequencies given by $n_{ij} / \sum_k n_{ik}$. All values shown in this example are entirely hypothetical and merely serve to illustrate the function of an allocation threshold when likelihood ratios are near unity. Figure 9.1-1 shows observed relative frequencies for the data from table 9 .1-1. The data exhibits some degree of overlap around $X = 5$. The continuous columns refer to Π_1 and the broken columns refer to Π_2.

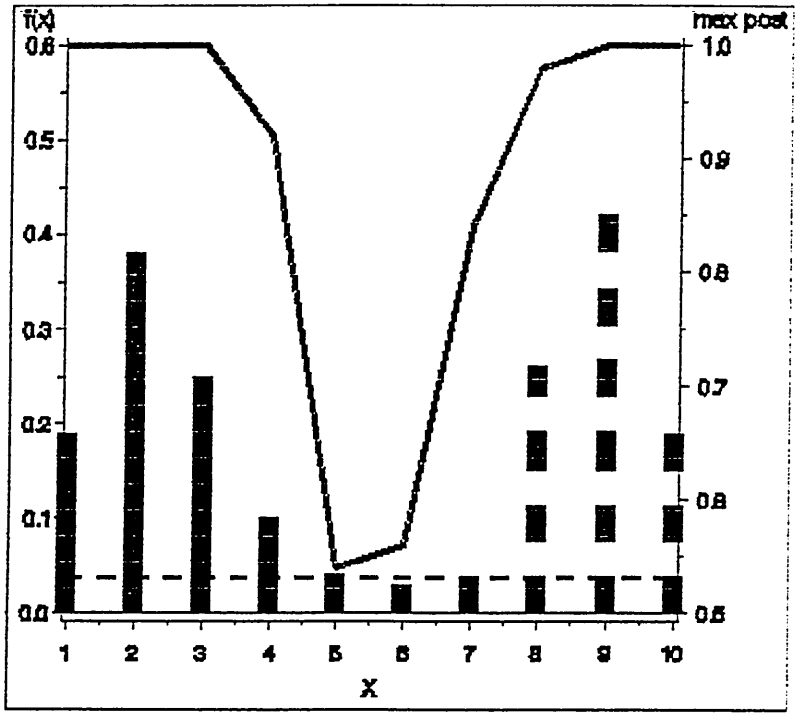


Figure 9.1-1: Thresholding min(h) at *0.53*

Figure 9.1-1 shows the effect of fixing the allocation threshold h^{min} at *0.53*. The maximum posterior $h^{(1)}$ is plotted against *X*. The left axis shows relative observed frequency, the right axis shows maximum posterior, which in the case of *2* populations varies between *0.5* and *1.0*. In the case of $h^{min} = 0.53$ this results in no rejections and corresponds to a single cutoff point. The misallocation error $\varepsilon^{counting}$, including rejections, is *0.08*.

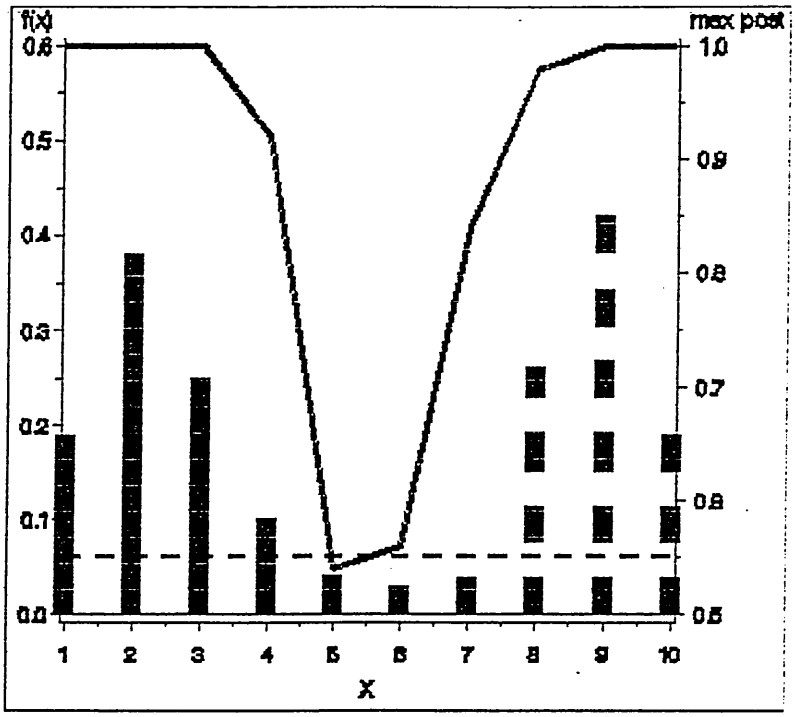


Figure 9.1-2: Thresholding min(h) at *0.55*

Figure 9.1-2 is the same as figure 9.1-1 but with $h^{min} = 0.55$. This now implies the cutoff region $\{x_{left}, x_{right}\}$ around $x = 5$ instead of a single cutoff point and stipulates that all objects with values of $x = 5$ are to be excluded because the maximum posterior $h^{(1)}$ for this value does not exceed *0.55*. If the rejected cases are treated as misallocations the misallocation error $\varepsilon^{counting}$ is increased to *0.12*.

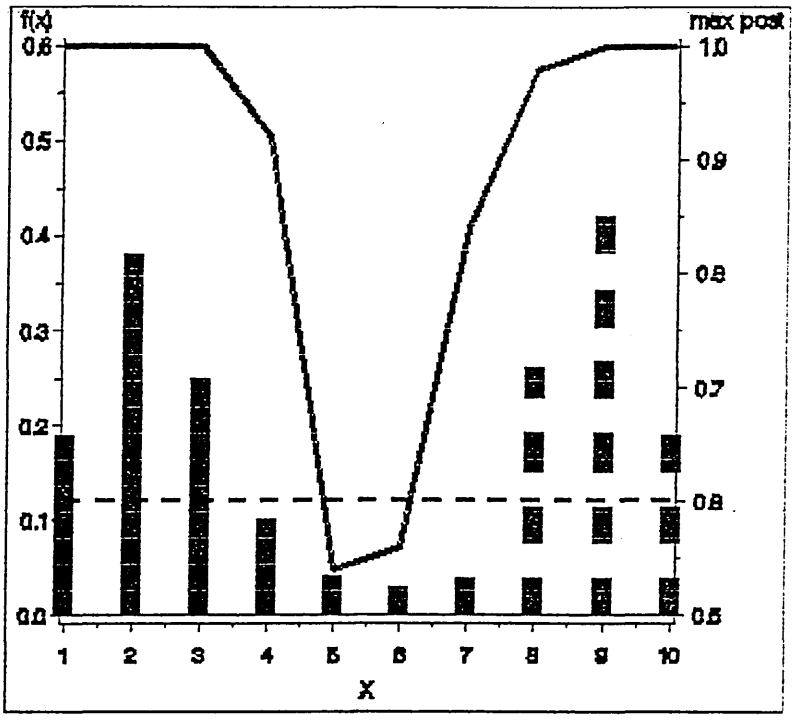


Figure 9.1-3: Thresholding min(h) at *0.60*

Figure 9.1-3 is again the same as figure 9.1-1 but now with h^{min} = *0.60*. This implies an even wider exclusion region $\{x_{left}, x_{right}\}$ embracing the values x = *5* and x = *6*. When rejections are treated as misallocations the error rate is increased to *0.15*. From figures 9.1-1 to 9.1-3 it can be seen that gradually increasing the minimum posterior probability h^{min} to be exceeded will result in a correspondingly increased proportion of rejected objects. If the rejected objects are treated as misallocations this of course results in an inflated error rate. In the present example the misallocation error rises from *0.08* for h^{min} = *0.53* via *0.12* for h^{min} = *0.55* to *0.15* for h^{min} = *0.60*.

Some statistical software packages for discriminant analysis allow the user to specify minimum posterior threshold values to be exceeded as a condition for allocation. Such *fixed allocation thresholds* allow allocations to population Π_i only if the posterior probability $f(\Pi_i|x)$ exceeds some function $\zeta(\tau,g)$ of the number of populations g and a scale factor τ allowed to

range between *0* and *1*. Let $\zeta(\tau, g)$ be the condition such that an object x_j whose true population membership is unknown is allocated to population Π_i ($i=1,..,g$) if

$$f(\Pi_i|x_j) > \frac{1+\tau}{g} \quad . \qquad (9.1\text{-}1)$$

The minimum posterior to be exceeded pertains when $\tau = 0$. This corresponds to the standard approach, where allocation simply depends on the maximum posterior $\max_i\{f(\Pi_i|x)\}$. Setting $\tau = 0.10$ means stipulating that posteriors have to be at least *10* percent greater than the minimum value of g^{-1}. This corresponds to setting a *fixed* threshold because reference is always to g^{-1} which is constant.

9.2 Thresholding and misallocation errors

As seen above raising a classification threshold will lead to inflated misallocation errors. The effect of varying the relative excess of $h^{(1)}$ over $h^{(2)}$ smoothly from *0* to *1* is demonstrated for two of the discrete datasets (*CESAR4*, and *CREDIT*) to be analysed further in chapters 13, 14 and 15. A detailed description of these datasets is given in chapter 11.

The *CESAR4* example set is a *2*-population, *4*-variate dichotomous medium sized dataset ($N = 1544$) concerned with the prediction of delivery by caesarean section from obstetric data gathered antenatally. Error rates and percentage classified were obtained after applying the linear discriminant procedure to the data. For the sake of demonstration the actual procedure selected is of secondary importance. Figure 9.2-1 shows misallocation error and proportion of allocated cases (due to not meeting the threshold criterion τ) plotted against τ. The proportion of allocated cases (continuous line) decreases with τ while the apparent error rate (broken line) increases.

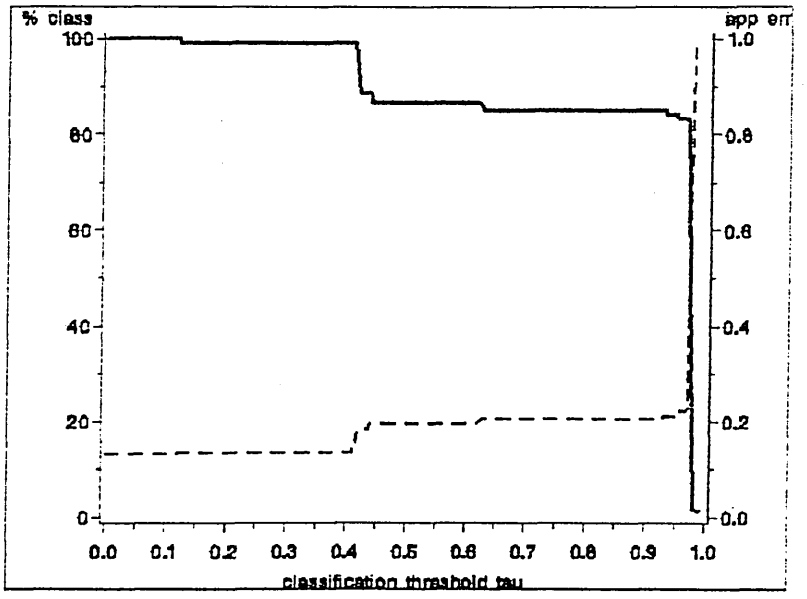


Figure 9.2-1: Effect of τ on error rate (*CESAR4*)

The *CREDIT* example is a *2*-population, *6*-variate ordinal medium sized (*N = 1000*) dataset concerned with the prediction of credit worthiness of bank customers on the basis of previous banking history and general sociodemographic data. Corresponding statistics are shown in figure 9.2-2. The proportion of allocated cases (continuous line) decreases with τ while the apparent error rate (broken line) increases.

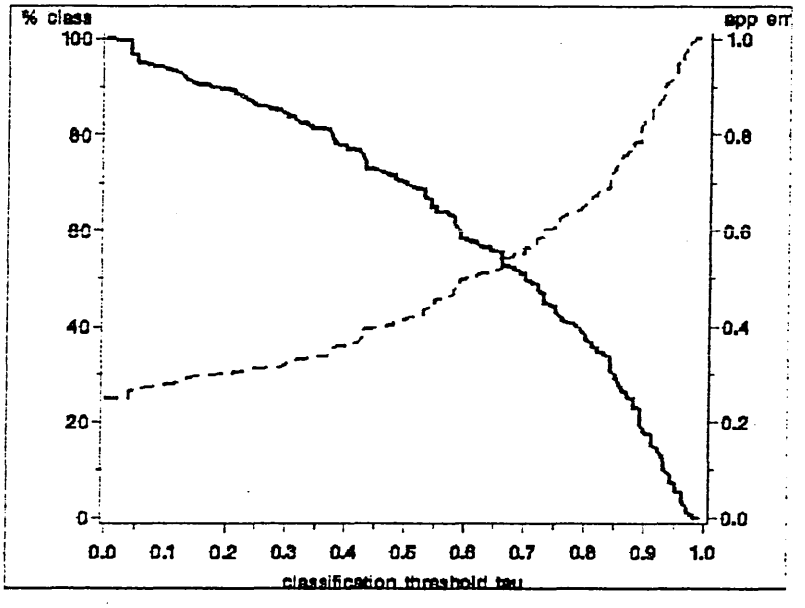


Figure 9.2-2: Effect of τ on error rate (*CREDIT*)

Both plots show clearly that a balance needs to be struck between choosing a threshold such that the error is minimised and selecting a threshold that minimises the proportion of rejected cases. The curves for the *CREDIT* data are smoother than those for the *CESAR4* data because the former dataset has more discrete states (see chapter 11) and thus exhibits smaller jumps in its density estimates.

9.3 Variable classification thresholds

A consequence of fixed thresholds is that functions of the form $\zeta(\tau, g)$ have the drawback that for $g > 2$ populations allocations will still be possible even if the two largest posteriors are similar in size. Consider a *3*-population situation where $h^{(1)} = 0.46$, $h^{(2)} = 0.44$ and $h^{(3)} = 0.10$. $h^{(1)}$ is well above the minimum threshold of $g^{-1} = 0.33$. Classically therefore allocation would be to Π_1 although $h^{(1)}$ only just exceeds $h^{(2)}$ thus casting doubt on the reliability of allocating to Π_1. As an illustration consider figure 9.3-1.

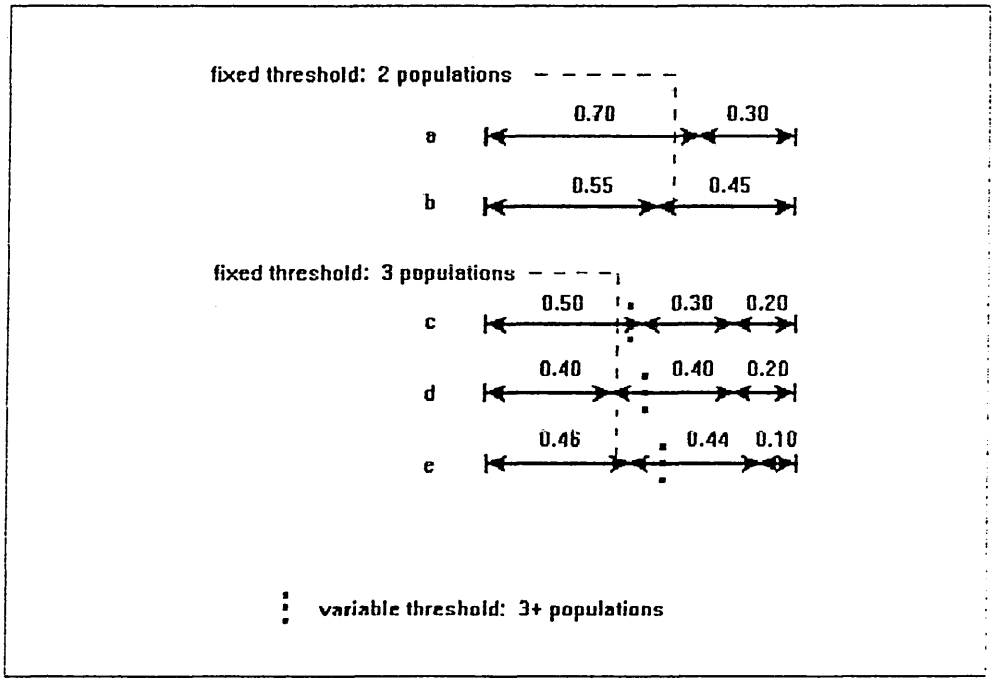


Figure 9.3-1: Limits of fixed thresholds

Fixed classification thresholds are indicated in figure 9.3-1 for the *2*-population examples a) and b) and the *3*-population examples c), d) and e) by the light dashed vertical lines. Variable thresholds are indicated in addition for the *3*-population examples by the heavy dashed vertical lines. For the *2*-population situation fixed and variable thresholds coincide. Posteriors for each example are drawn horizontally with all lines summing to unity. The use of fixed thresholds in example e) leads to the paradoxical situation that allocation is made to population Π_1 although the corresponding largest posterior $h^{(1)} = 0.46$ is only marginally bigger than the second largest posterior $h^{(2)} = 0.44$. This paradox can however be overcome by using *variable classification thresholds*. Stipulating a minimum relative difference between the two largest posteriors $h^{(1)}$ and $h^{(2)}$ is therefore suggested instead as a more adequate filter. The rule now becomes for the first population: *Allocate a given object* x_0 *to population* Π_1 *if*

$$\frac{h^{(1)} - h^{(2)}}{h^{(1)}} > \tau \qquad (9.3\text{-}1)$$

where $0 \leq \tau \leq 1$. Assume g populations and for any object $\mathbf{x}_j$ let $h^{(1)}, h^{(2)}, \ldots, h^{(g)}$ be the sequence of *order statistics* (Lindgren, 1976) on the posteriors h_i such that $h^{(1)} \geq h^{(2)} \geq \ldots \geq h^{(g)}$. Then expression (9.3-1) is equivalent to

$$1 - \frac{h^{(2)}}{h^{(1)}} > \tau \qquad (9.3\text{-}2)$$

or

$$h^{(1)} > \zeta(\tau, h^{(2)}) \ .$$

It follows from (9.3-2) that the threshold for the maximum posterior $h^{(1)}$ will only vary depending on the size of $h^{(2)}$.

The superiority of variable classification thresholds over fixed ones can be demonstrated theoretically by considering

the general case of discrimination between g populations. The fixed threshold is constant and lies above g^{-1}. For any given observation **x** two extreme conditions may be distinguished for the distribution of the posteriors: Case(1): $h^{(2)}$ is at a minimum and all posteriors less than the maximum posterior are at most equal ($h^{(2)} = h^{(3)} = \ldots = h^{(g)}$) whence $h^{(1)} + (g-1)\, h^{(2)} = 1$. Case(2): $h^{(2)}$ is at a maximum and all posteriors beyond the second largest are zero ($h^{(3)} = h^{(4)} = \ldots = h^{(g)} = 0$) whence $h^{(1)} + h^{(2)} = 1$. All possible values for the relative difference $\tau = (h^{(1)} - h^{(2)}) / h^{(1)}$ will lie on an interval whose limits are given by these two conditions. When $h^{(2)} = h^{(3)} = \ldots = h^{(g)}$ (case(1)) it follows from

$$h^{(1)} + (g - 1)\, h^{(2)} = 1 \qquad (9.3\text{-}3)$$

that

$$h^{(2)} = \frac{1 - h^{(1)}}{g - 1} \qquad (9.3\text{-}4)$$

and as

$$\tau = \frac{h^{(1)} - h^{(2)}}{h^{(1)}} \qquad (9.3\text{-}5)$$

upon substitution of $h^{(2)}$ from 9.3-4 into 9.3-5 τ becomes

$$\tau = \frac{h^{(1)} - \dfrac{1 - h^{(1)}}{g - 1}}{h^{(1)}} \qquad (9.3\text{-}6)$$

whence

$$h^{(1)} = \frac{1}{g + \tau(1 - g)} \,. \qquad (9.3\text{-}7)$$

The expression for $h^{(1)}$ in 9.3-7 constitutes a lower bound for $h^{(1)}$ under a variable classification threshold scheme. By a similar argument when $h^{(3)} = h^{(4)} = \ldots = h^{(g)} = 0$ (case(2)) an upper bound for the threshold posterior may be shown to be given by

$$h^{(1)} = \frac{1}{2 - \tau} \; . \qquad (9.3\text{-}8)$$

The behaviour of the upper and lower threshold posterior bounds $h^{(1)}$ given in expressions 9.3-7 and 9.3-8 is sketched in figures 9.3-2 to 9.3-4 for values of $g \in \{3,4,6\}$ against relative differences between the two largest posteriors τ in the range *0* to *1*. In addition the fixed threshold curve $h^{(1)} = (1 + \tau) \; / \; g$ is also drawn.

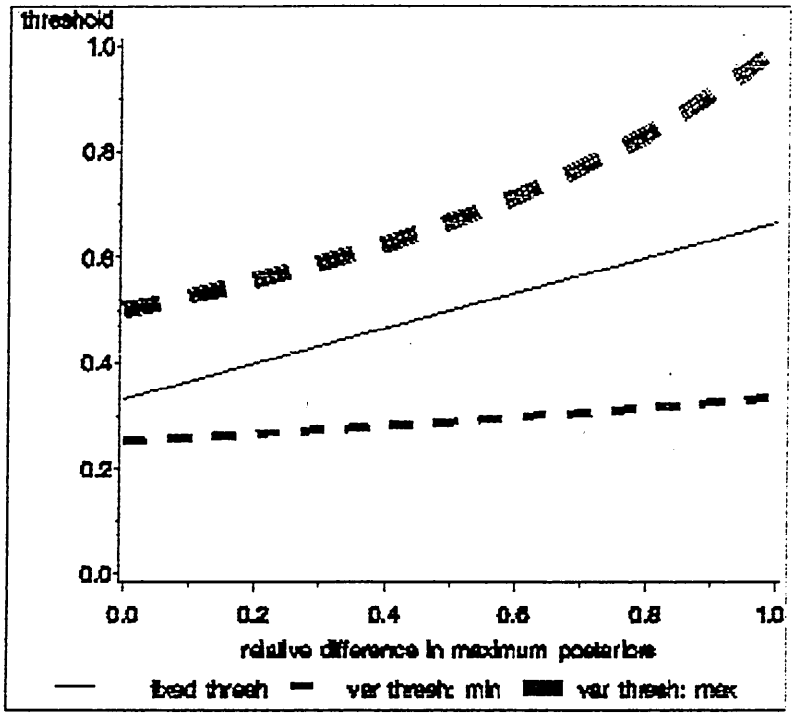


Figure 9.3-2: Threshold bounds for *3* groups

A comparison with the corresponding fixed threshold shows that for *3* populations generally the fixed threshold lies about half way between the lower and upper bounds (9.3-7 and 9.3-8). As the number of populations increases the fixed threshold moves gradually towards the lower bound for the variable threshold. This means that for *4* and more populations the value of the variable classification threshold lies mainly in preventing *marginal* allocations where $h^{(1)}$ is only just slightly larger than $h^{(2)}$. This however is just the typical paradox situation that it was intended to deal with by using variable as opposed to fixed thresholds.

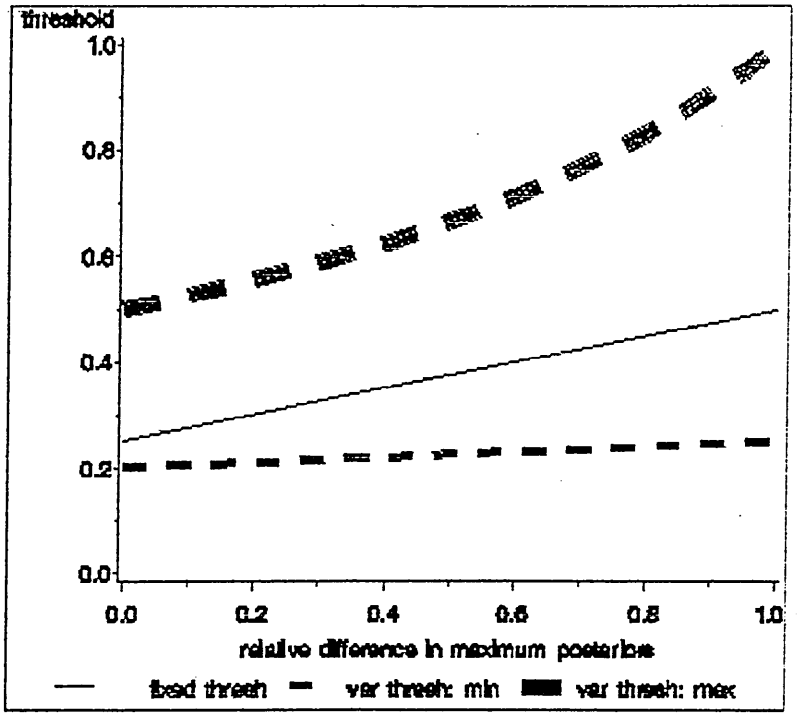


Figure 9.3-3: Threshold bounds for *4* groups

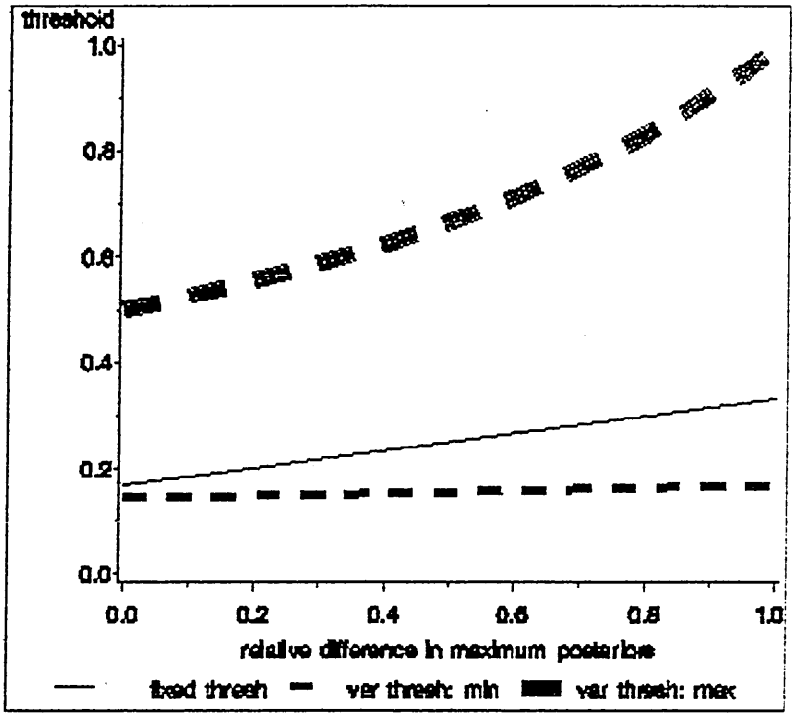


Figure 9.3-4: Threshold bounds for *6* groups

With increasing number of populations g the difference between the fixed threshold and the lower variable limit decreases. Two situations may be distinguished:

(a) The region bounded by the continuous line and the lower broken line indicates such cases where use of variable thresholds might lead to additional allocations that a fixed threshold would not accept. Here potential is seen for variable thresholds to lead to increasing the percentage of classified objects and thus to lowering error rates.

(b) The region bounded by the continuous line and the heavy broken upper line indicates cases where use of the fixed threshold would lead to allocations to Π_1 although the corresponding posterior $h^{(1)}$ only just exceeds $h^{(2)}$ thus resulting in allocations with smaller reliability. With increasing number of populations the probability of such *marginal* allocations rises under the fixed threshold scheme. Use of variable thresholds would prevent this.

9.4 Assessing performance from $f(\tau)$

Another way of describing the relationship between performance of a discriminant and the distribution of the largest posteriors $h^{(1)} = \max_i\{h_i\}$ is in terms of the distribution of relative differences between the two largest posteriors. Let $\tau = (h^{(1)} - h^{(2)}) / h^{(1)}$ such that τ is bounded: $0 \leq \tau \leq 1$. Datasets with little overlap between populations will exhibit distributions for τ that are skewed towards *1*. This feature may be illustrated with the dataset *B* from chapter 8 where the misallocation error $\varepsilon_{\text{counting}}$ is comparatively low (*0.040*) and thus indicative of a small degree of overlap between Π_1 and Π_2.

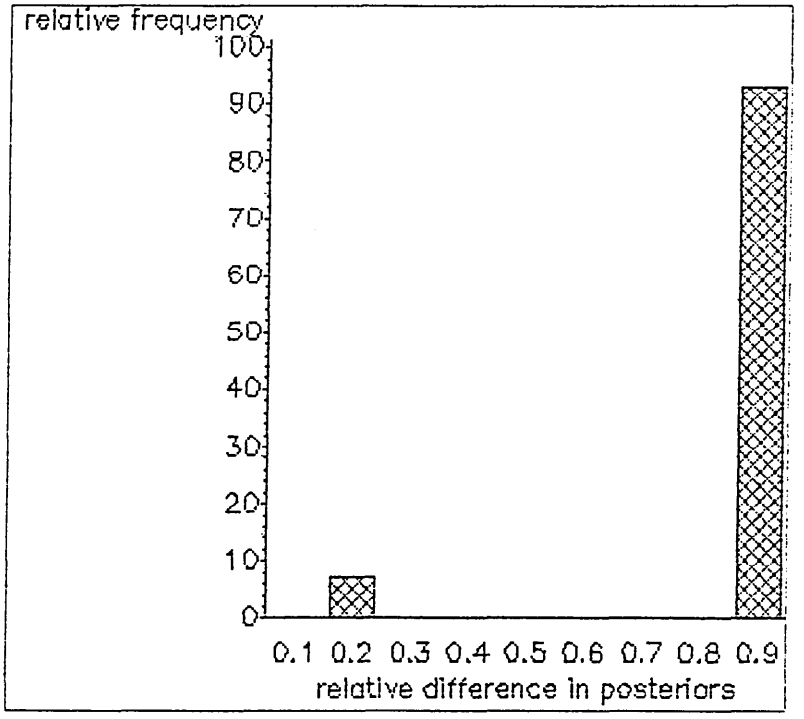


Figure 9.4-1: $f(\tau)$ distribution for dataset *B*

Figure 9.4-1 shows a histogram of the corresponding distribution of relative differences τ which exhibits a strong positive skew. Values of τ are plotted along the horizontal axis ranging from *0* to *1*. The distribution itself is bimodal with the majority of relative differences lying in the upper range near *1*. This fact is indicative of good separability. The misallocation error of *0.040* is correspondingly low.

Datasets in which the degree of overlap between different populations is large generally yield discriminant functions showing a considerable proportion of posteriors of similar size where the relative differences τ are smaller.

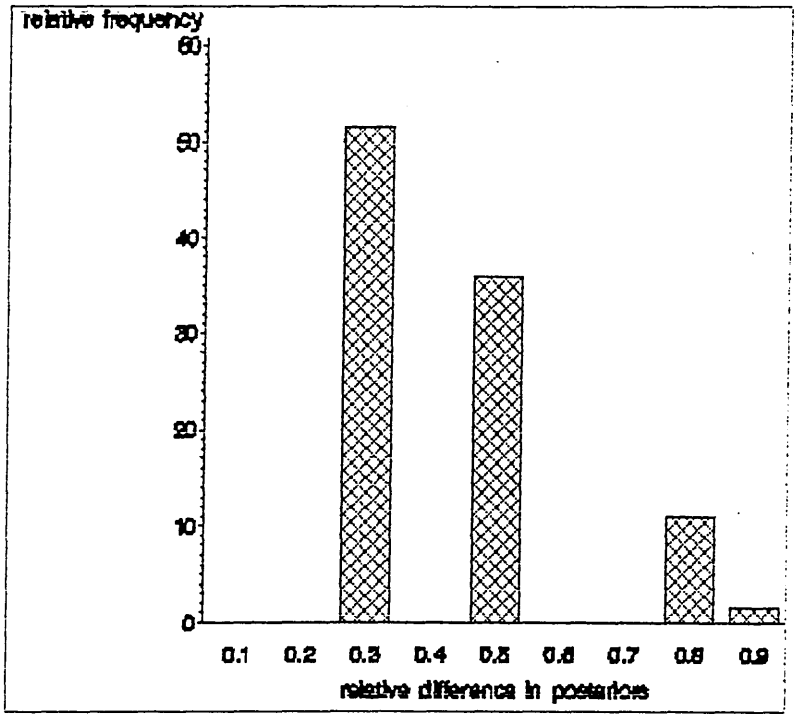


Figure 9.4-2: $f(\tau)$ distribution for dataset *A*

Figure 9.4-2 depicts this situation for the dataset *A* from chapter 8 where $\varepsilon_{counting}$ was *0.275*. In contrast to dataset B the distribution shows the majority of relative differences lie below values of *0.60*. This confirms the high misallocation error. The distribution of τ is negatively skewed. As was seen in section 9.2 the proportion of rejected objects is a function of the allocation threshold τ. If rejections are treated as misallocations the error rate will generally increase along with τ. The exact nature of the relationship between ε and τ will however depend on the sampling distribution of $\tau = (h^{(1)} - h^{(2)}) / h^{(1)}$. If, for instance the weight of the distribution of τ lies skewed towards *1* then ε will only change slightly as τ increases away from *0*. Conversely a negatively skewed τ distribution with considerable mass near *0* will lead to rapid increases in $\varepsilon_{counting}$ even for moderate changes in τ. Two diagrams may help to illustrate this point:

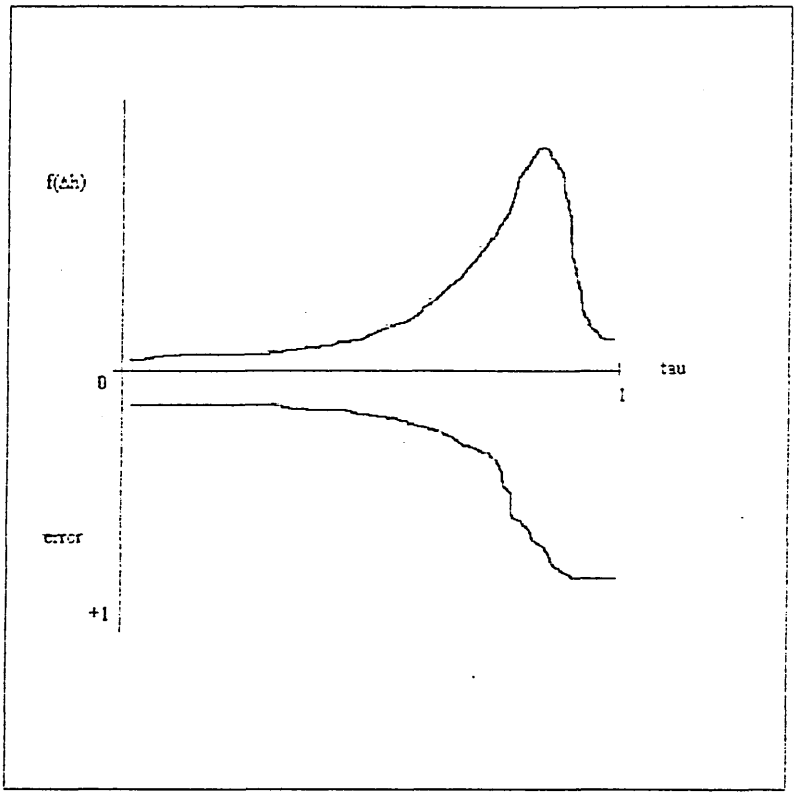


Figure 9.4-3: Positively skewed τ distribution

Figure 9.4-3 shows an illustrative drawing of the hypothetical course of the misallocation error as a function of relative difference in posteriors τ. The behaviour of the error rate depends on the distribution of τ. In the above case the distribution of τ is skewed towards *1*. The majority of posteriors show large relative differences suggestive of good separation. The corresponding error rate thus rises[25] only slowly with initial values of τ and picks up substantially only at much higher threshold levels.

[25] (n.b.: The value of *+1* for the misallocation error has been plotted downwards)

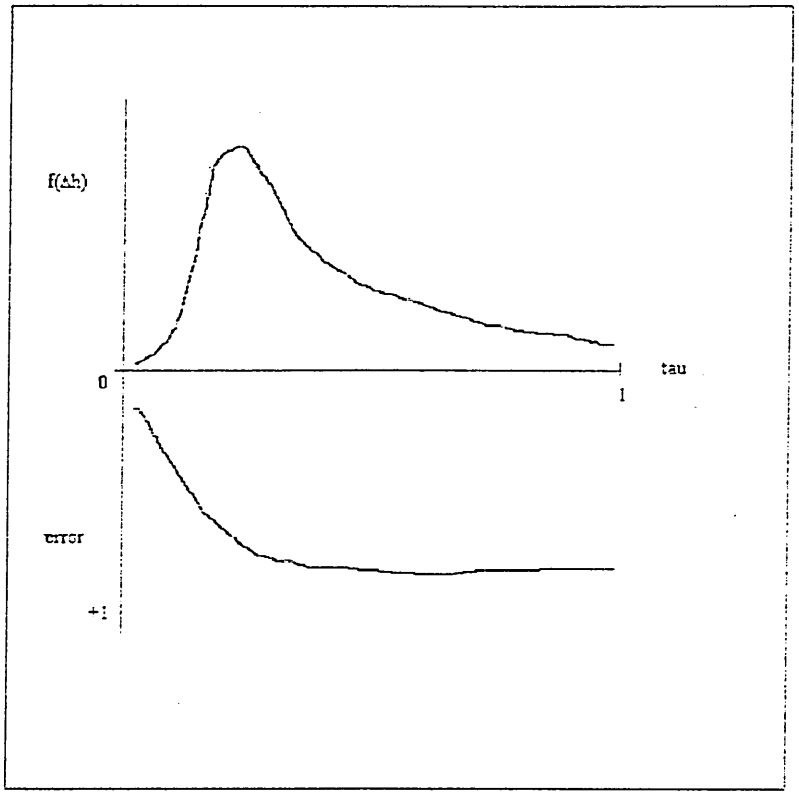


Figure 9.4-4: Negatively skewed τ distribution

Figure 9.4-4 is another illustrative drawing and similar to figure 9.4-3. However, in the this case the distribution of τ is skewed towards *0*. The majority of posteriors show small relative differences suggestive of poor separation. Here smaller changes in τ will lead to initially dramatic increases in the error rate. Later the error rate levels off as the density $f(\tau)$ decreases.

Figures 9.4-3 and 9.4-4 show hypothetical examples of clearly different distributions yet for $\tau = 0$ the misallocation errors are similar. Commonly in applications of discriminant analysis non-thresholded misallocation error rates are quoted corresponding to $\tau = 0$. There is nothing wrong in quoting error rates at *zero* threshold levels $\varepsilon_{\tau=0}$ as long as the quote comes along with some measure of confidence. Clearly sampling effects will be much greater on estimates of misallocation errors at zero or in the region of zero ($\varepsilon_{\tau=0}$ and $\varepsilon_{\tau\sim0}$) in the situation given in figure 9.4-4 than in figure 9.4-3. Of course these two situations are theoretical because the selected

distributions are deliberately chosen to be extreme. As the absolute value of ε is correlated with the shape of the distribution of τ the misallocation error will tend to be small when the distribution of τ is positively skewed and vice versa.

Several approaches of utilising the $f(\tau)$ distributions for assessing the performance of discriminants are possible:

(1) A first approach focuses on the shape of the estimated curves for $f(\tau)$. In particular their degree of positive skewness is seen as the basic entry point for further qualifying estimates of performance criteria[26]. Here a visual inspection though not very formal but informative appears most useful. Good discriminants would exhibit distributions similar to those sketched in figure 9.4-3.

(2) A second approach consists of formalising (1). A measure of reliability of a discriminant is constructed to reflect the rate at which the empirical threshold distribution increases per unit interval along some selected subsection of the τ axis. Here the focus might be on the initial part of the axis where thresholds lie in a range ($0 \leq \tau \leq \alpha$). Good performance would then be indicated by low rates of increase within this interval. The second approach is essentially the same as (1) but places more emphasis on the lower end of the τ range. This is more sensitive to *marginal* allocations where $h^{(1)}$ is only slightly larger than $h^{(2)}$.

(3) When classification thresholds are used this is of consequence for the error rate. Thus the misallocation error ε is a function of τ. A discriminant function with good separating ability will be characterised by a positively skewed τ distribution with initially slow rise in $f(\tau)$ such as given in figure 9.4-3. The corresponding

---

[26] This will become clear in chapter 14.

error rate ε will also rise more slowly in this case. Consider the illustrative drawing shown in figure 9.4-5.

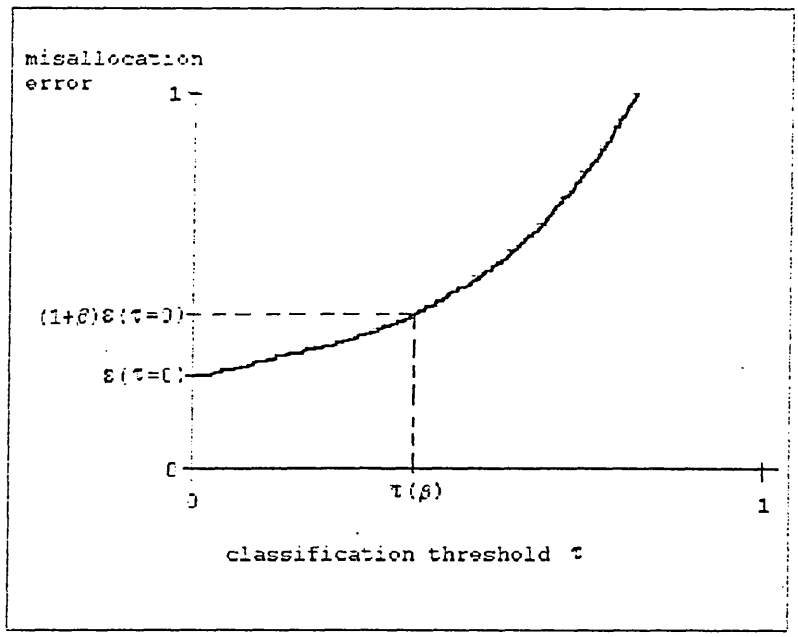


Figure 9.4-5: Threshold dependent error rate

Initially when $\tau = 0$ the corresponding error rate $\varepsilon(\tau=0)$ at this point is also at a minimum. Let $\tau(\beta)$ be the value of τ for which the initial error rate $\varepsilon(\tau=0)$ is increased by a proportion β to $(1 + \beta)\ \varepsilon(\tau=0)$. The further $\tau(\beta)$ is away from 0 the slower the rise in ε and the better the performance of a given discriminant function yielding the underlying $f(\tau)$ distribution. The usefulness of the above approach is inspected briefly in chapter 14 for the case of $\beta = 1$ thus leading to what may be called *error doubling points*.

(4) A further formalisation of (3) is given when $\hat{f}(\tau)$ is used as a general measure for any performance criterion. Let $\varphi(.)$ be a given performance criterion such as $\varepsilon_{counting}$, $\varepsilon_{posterior}$ or η. Then a new criterion φ^* may be derived as

$$\varphi^* = \int_0^1 \varphi(.)\ \hat{f}(\tau)\ d(\tau)\ . \qquad (9.4\text{-}1)$$

Use of expression 9.4-1 has so far not been investigated further but it is suspected that φ^* and $\varepsilon_{\tau=0}^{counting}$ will be strongly correlated. In discriminant problems the nature of the given data and the aims for which the discriminant is constructed will frequently determine how performance criteria are valued. Thus before deciding on a particular formalised rule for interpreting $f(\tau)$ it is probably initially sufficient just to inspect the actual resultant performance curves (approach 1 above) and interpret these set against demands placed on the particular discriminant problem[27].

9.5 Summary

Classification thresholds are introduced for two purposes: (a) to control the reliability of an allocation rule estimated from a discriminant procedure and (b) to facilitate selection of a discriminant procedure. Purpose (a) is achieved by reduction of *marginal allocations* by specifying a minimum threshold, τ_{min}, and purpose (b) is achieved by generating empirical distributions of relative differences, $f(\tau)$, between the two largest posteriors and inspection of threshold dependent performance, $\varphi(\tau)$.

Considerable use could be made by the introduction of such variable classification thresholds for $g > 2$ populations as popular statistical software packages for discriminant analysis currently only offer the option of specifying a constant threshold.

Classification thresholds have their price as rejected allocations inflate the error rates. Consequences of this feature are illustrated for two real datasets.

The application of a fixed threshold to discrimination among $g > 2$ populations is shown to not fully exploit the

---

[27] see also chapter 12

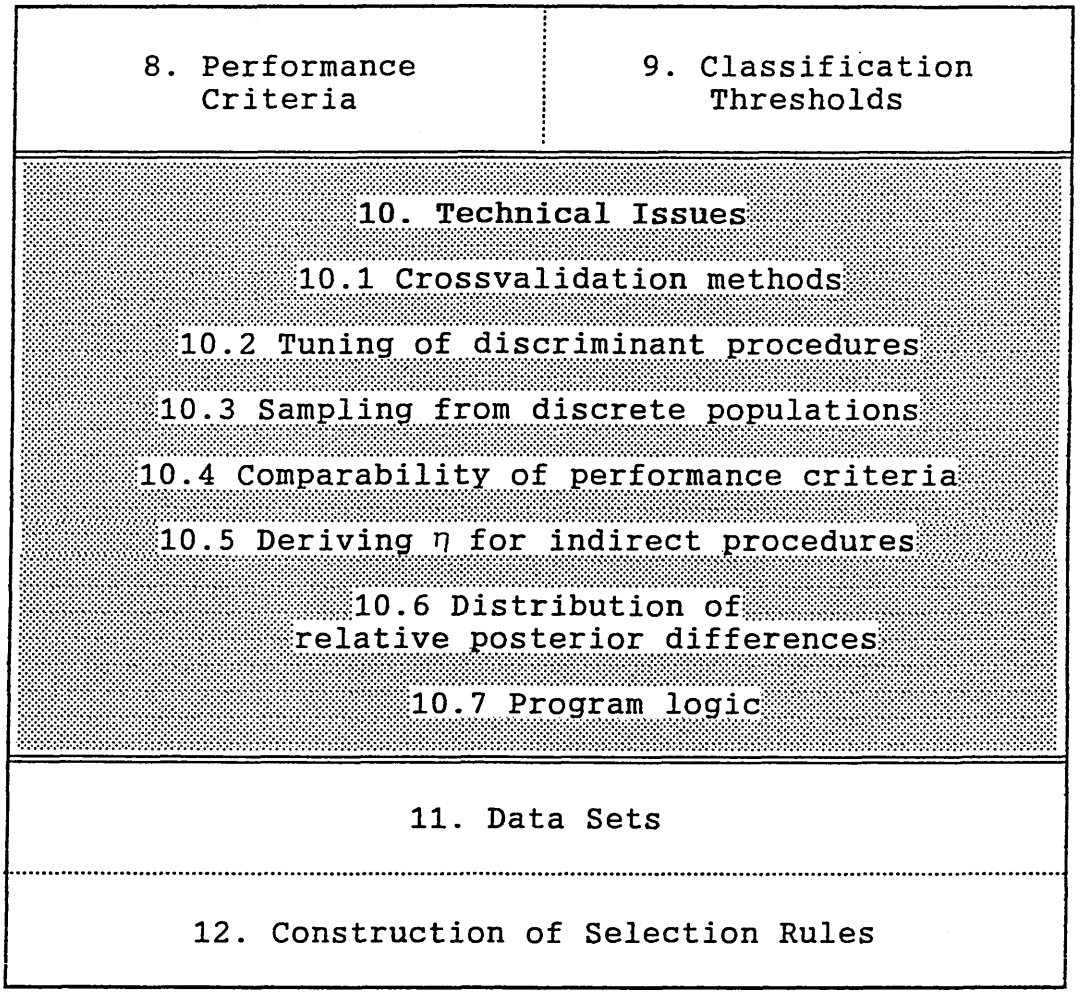
information contained in the posteriors. In the case of $g = 3$ populations, for instance, the relative difference between the two largest posteriors, $h^{(1)}$ and $h^{(2)}$, may vary considerably. In order to compensate for this *variable classification thresholds* are introduced. This concept will be used later when analysing datasets with $g \geq 3$ populations.

Two tools based on variable classification thresholds are suggested as aids to help selection of a discriminant procedure: (1) the empirical distribution, $f(\tau)$, of relative differences between the two largest posteriors for a given discriminant procedure and (2) threshold dependent plots of performance criteria, $\varphi(\tau)$, as a function of classification threshold, τ. Both these tools will be used in chapter 12 to construct the *selection tree* and also in chapters 14 and 15 in actual applications to real and artificial datasets.

I: INTRODUCTION

II: REVIEW

III: METHOD



IV: RESULTS

V: DISCUSSION

Before executing the various discriminant procedures for discrete data and computing the performance criteria under different posterior thresholding conditions, some adjustments are necessary in order to ensure that comparable and meaningful results can be generated. In chapters 7, 8 and 9 basic notions of performance evaluation and crossvalidation were introduced. In the following sections technical details of the actual operation of the discriminant procedures and crossvalidation techniques are described. Section 10.1 discusses application of the crossvalidation and bootstrap techniques. In section 10.2 the fine tuning of the kernel density estimation based discriminant procedure and of the Hills distance based procedure is described. Section 10.3 addresses the problem of sampling from discrete distributions with sparse states. Section 10.4 gives final adjustments made to $\varepsilon^{counting}$, $\varepsilon^{posterior}$ and η to facilitate comparison of bias and variance between all three performance criteria. Section 10.5 explains how the performance criteria are evaluated differently for indirect and indirect procedures. Section 10.6 outlines the metric for generating the distribution $f(\tau)$ of relative differences in the *2* largest posteriors. The logic of the program structure is explained in section 10.7.

10.1 Crossvalidation methods

All four crossvalidation techniques discussed in chapter 9 are used: *resubstitution*, *hold-out* with equally sized training and test sets, *leave-one-out* (Lachenbruch, 1975) and *leave-v-out* crossvalidation with *v* set to *10%*. The parameter *v* was set to *10%* because this places the *leave-v-out* technique roughly between the *leave-one-out* and *hold-out* techniques. By doing this it is hoped that the greater proximity to the *leave-one-out* technique will make it

almost as effective as the *leave-one-out* while reducing the number of iterations per estimation cycle by a tenth.

10.1.1 Leaving-one-out crossvalidation

Leaving-one-out crossvalidation takes a special form in the case of discrete data and leads to simple computation formulae. Several observations typically share identical combinations of variables when data are discrete. Once the state probabilities $\{p_{ij}\}$ have been determined for a particular dataset these will also imply counts $n_{ij} = p_{ij}n_i$. On each leave-one-out cycle these counts are reduced by *1* giving a new cell count $n'_{ij} = n_{ij} - 1$ with $i=1,\ldots,g$ and $j=1,\ldots,s$. This leads to a new set of state probabilities $\{p'_{ij}\}$. Thus, for any given dataset only *sg* (= number of states times number of populations) crossvalidation cycles are required which is a number generally far less than the actual number of objects *n* in the sample. This fact also justifies on economical grounds the use of crossvalidation and bootstrap techniques for discrete data situations, a feature particularly useful when computing conditional and unconditional performance estimates (see section 10.1.2).

In chapter 8 the expression for $\varepsilon^{\text{counting}}$ is given in terms of an individual observation $\mathbf{x}_j$ and the indicator function $\gamma_i(\mathbf{x}_j)$. In the state matrix notation this expression simplifies further. Recall that $\gamma_i(\mathbf{x}_j) = 1$ whenever $\mathbf{x}_j \in \Pi_i$ and zero otherwise. To pick out the appropriately weighted proportion of misallocations, given that p'_{ij} now represents the respective state probability in the test sample, the following expression is used for $\varepsilon^{\text{counting}}$:

$$\varepsilon^{\text{counting}}_{\text{crossval}} = \sum_{i=1}^{g} \pi_i \sum_{j=1}^{s} p'_{ij} \, (1 - \gamma_i(\mathbf{x}_j)) \qquad (10.1\text{-}1)$$

For the other crossvalidation techniques, hold-out and leave-v-out, corresponding formulae were used in an analogous fashion. Similar expressions were used also for $\varepsilon^{posterior}$ and η.

10.1.2 Bootstrapping

The following stages describe the *nonparametric bootstrap* technique (Efron, 1979) used to estimate population statistics. It has been shown by McLachlan (1980) and Schervish (1981) to be highly efficient for this purpose when compared to the *parametric bootstrap*. For this reason it was employed here. The bootstrap method was used for two purposes: to estimate the bias of the performance criteria, and to estimate the distribution of *relative differences* between the two largest posterior probabilities. Under the correct model assumptions, the relative difference between the two largest posteriors is assumed to be directly related to the certainty with which an observation should be allocated to one of g populations (chapter 9).

Bootstrap techniques were applied to obtain estimates of the variance and bias of the performance criteria. Two cases are distinguished: *conditional* and *unconditional* performance (chapter 7).

Consider first *conditional* performance: Given a training dataset $\mathbf{t}$[28] a discriminant rule $\delta(\mathbf{x};\mathbf{t})$ is initially derived using a chosen discriminant procedure. Next the r^{th} ($r=1,\ldots,R$) *bootstrap replicate*, $\mathbf{t}_r^*$, is generated from $\mathbf{t}$. This is achieved by separate sampling (section 10.3) of the observed empirical cumulative distribution. The r^{th} bootstrap estimate of the discriminant's *conditional* performance, φc_r^*, is formed by applying the allocation rule $\delta(\mathbf{x};\mathbf{t})$ to the r^{th} bootstrap replicate, $\mathbf{t}_r^*$. This process is repeated for $r = 1,\ldots,R$ where R is the total number of

[28] The bold letter $\mathbf{t}$ denotes a sample, not a vector.

bootstrap *replicates*. The average over R such replicates constitutes a bootstrap *trial* which provides an estimate of the *conditional performance*[29] $\varphi c^*(\delta|\mathbf{t})$. This is the performance of a rule δ estimated from the original sample $\mathbf{t}$ and applied R times to bootstrap samples $\mathbf{t}_1^*,\ldots,\mathbf{t}_r^*,\ldots,\mathbf{t}_R^*$. The process consists of seven steps: Step (1) initialises the population specific *empirical distribution function*, $\hat{F}_i$, used for bootstrap sampling. Steps (1) to (7) constitute one bootstrap *trial* from which estimates of *conditional performance* are gained.

(1) Compute the *empirical cumulative distribution functions*, $\hat{F}_i(\mathbf{x})$, of q-variate discrete observations, $\mathbf{X}$, from the training data separately for each population. The observed relative frequencies of occurrences of combinations of values of X_{ik} $(k=1,\ldots,q)$ may be represented in *state matrix notation*. This yields the multinomially distributed univariate variable $X^{(s)}\sim\mathfrak{M}(p_{i1},p_{i2},\ldots,p_{is})$ where the parameters $\{p_{ij}\}$ correspond to the vector of observed relative frequencies of observations $\mathbf{X}$.

(2) Compute the allocation rule $\delta(\mathbf{x}^{(s)}|\mathbf{t})$.

(3) Generate an observation $X^{(s)}$ from this distribution using a suitable random number generator.

(4) Repeat step (3) n_i times to generate a sample of size n_i for each of g populations giving the r^{th} bootstrap *replicate*, $\mathbf{t}_r^*$.

(5) Apply the allocation rule $\delta(\mathbf{x}^{(s)}|\mathbf{t}_r^*) = \delta(\mathbf{t}_r^*)$ to this replicate, compute and store the relevant performance statistics.

(6) Repeat steps (3) to (5) R number of times.

(7) Average the performance statistics generated in step (5) over the R replicates.

---

[29] The lower case $\mathbf{t}$ indicates that the expectation is with respect to the R bootstrap replicates.

Consider next *unconditional* performance: An estimate of *unconditional performance* may be obtained by repeating the above trial itself $t=1,\ldots,T$ times. This time however the allocation rule itself is one of T bootstrap estimates.

(1) Generate a bootstrap sample t_t^* for the t^{th} *trial*.

(2) Use this to derive the t^{th} bootstrap allocation rule $\delta_t^*(\mathbf{x}^{(s)}|t)$.

(3) For the t^{th} *trial* generate the r^{th} bootstrap *replicate*, t_{tr}^*, following the steps above.

(4) Apply $\delta_t^*(\mathbf{x}^{(s)}|t_{tr}^*) = \delta_t^*(t_{tr}^*)$, compute and store relevant performance statistics.

(5) Repeat steps (3) and (4) R times.

(6) Average the results from step (4) over R *replicates* yielding *conditional* performance for the t^{th} *trial*.

(7) Repeat steps (1) to (6) T times and average the *conditional* performance estimates from step (6) over T *trials* yielding *unconditional* performance.

The final average from step (7) above may be written as[30] $\varphi u^*(\delta|\mathbf{T})$. In this case the rule δ itself is estimated T times from independently generated bootstrap samples. It follows that estimates of unconditional performance are more computer intensive to derive as they involve $R * T$ estimation cycles.

In order to apply the bootstrap technique it was necessary to obtain an estimate of how many bootstrap trials T are required for achieving stable estimates of the estimated allocation rule, $\hat{\delta}$. It was decided to generate a sufficient number of allocation rules for various data types and discriminant procedures and to compute their variability using a root mean square measure. The number of replicates

[30] The upper case $\mathbf{T}$ indicates that now the expectation is taken with respect to the T bootstrap trials.

R on any one trial was fixed from the start at 100 following Efron (1979, 1990) and McLachlan (1992) who found that this is sufficient for (conditional) bias and variance estimates. Pilot experiments were next conducted on several discrete datasets of varying complexity in terms of number of discrete states $s = \sum_{k=1}^{q} l_k$ and varying cell proportions p_j, $j=1,\ldots,s$. First the expected allocations were estimated by taking expectations over $T=1000$ bootstrap trials for each of s discrete states

$$\bar{\delta}^*(j) = E_t\left[\delta_t^*(j)\right] \qquad (10.1\text{-}2)$$

with $j=1,\ldots,s$. The subscript for E_t in 10.1-2 indicates that the expectation is taken with respect to the T trials. The expected allocations, $\bar{\delta}^*(j)$, take on the integer values $i \in \{1,2,\ldots,g\}$. The root mean square measure is then given as a function of the number of bootstrap trials, t, by

$$RMS_t(\delta^*) = \left[\frac{1}{st} \sum_{v=1}^{t} \sum_{j=1}^{s} \left(\delta_v^*(j) - \bar{\delta}^*(j)\right)^2 \right]^{1/2}. \qquad (10.1\text{-}3)$$

The following figures 10.1-1 to 10.1-4 show the behaviour of $RMS_t(\delta^*)$ as defined in expression 10.1-3.

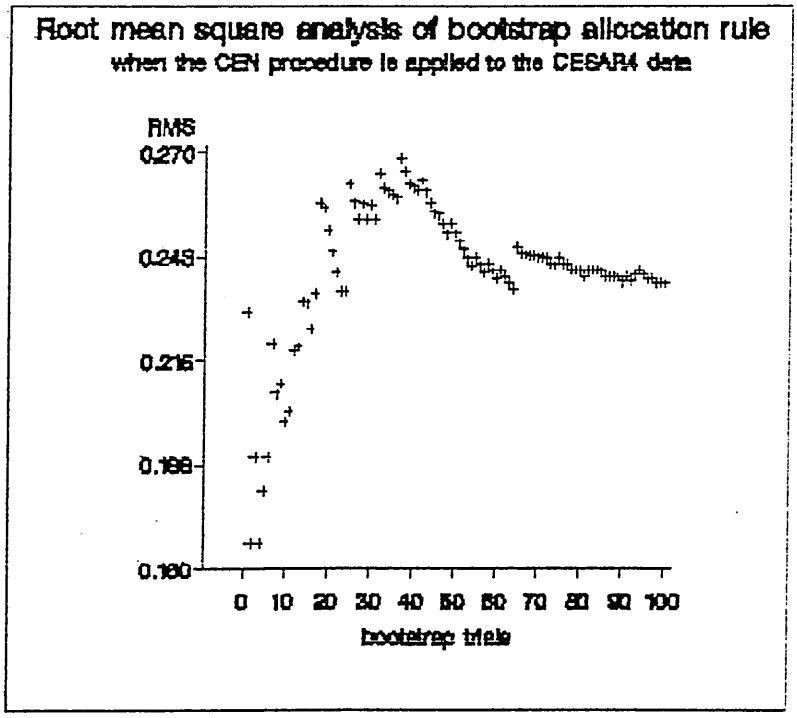


Figure 10.1-1: *RMS* analysis for *CESAR4* data.

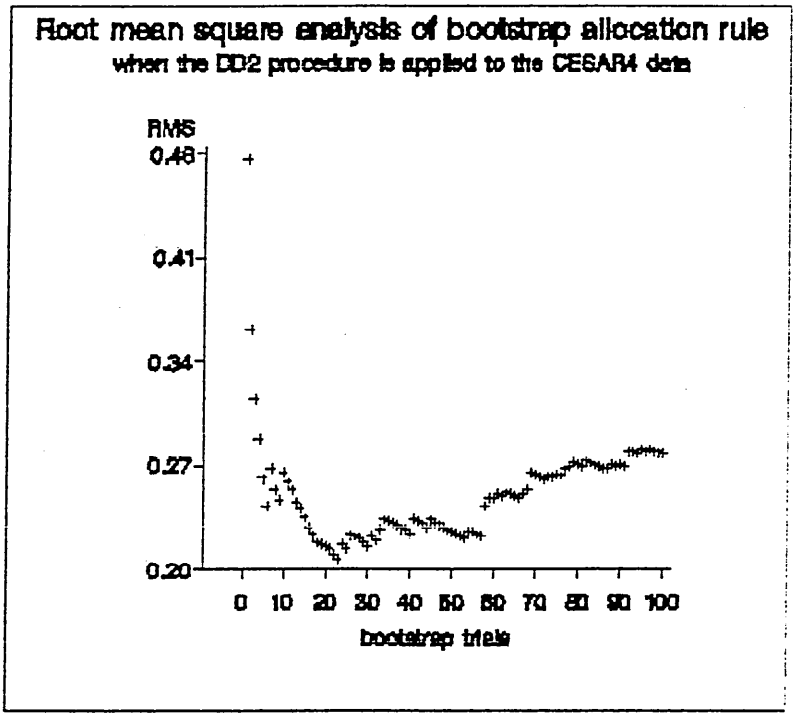


Figure 10.1-2: *RMS* analysis for *CESAR4* data

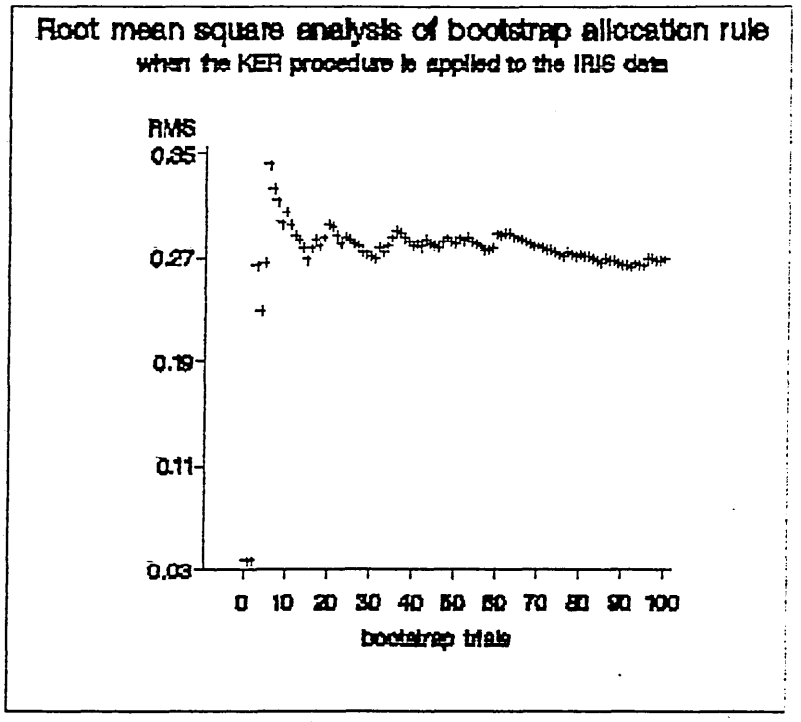


Figure 10.1-3: *RMS* analysis for *IRIS* data

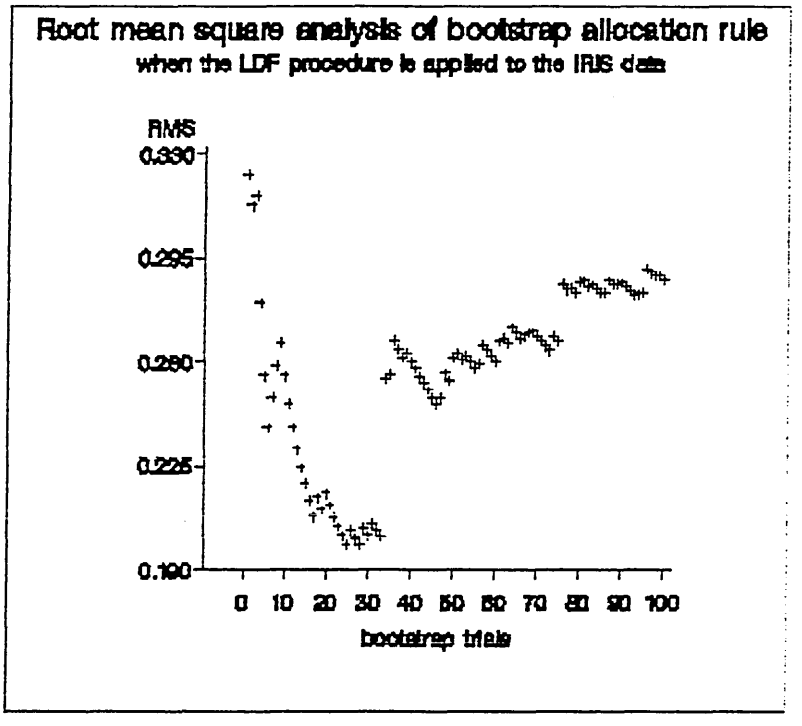


Figure 10.1-4: *RMS* analysis for *IRIS* data

Analyses are presented for the centroid (*CEN*) and the modified distributional distance model (*DD2*) applied to the

CESAR4[31] dataset and for the kernel (*KER*) and the linear discriminant (*LDF*) procedure applied to the *IRIS* dataset. Inspection of these figures shows that the variability of the bootstrap estimates of allocation rules tends to settle down at approximately *100* trials. Further analyses not shown here exhibit similar behaviour for $RMS_t(\delta^*)$. It was therefore decided to set $T = 100$ throughout.

Two independent runs of the sampling experiment were conducted. In the first expectations and variances of estimates φc of conditional performance of the discriminants were obtained from one trial consisting of R replicates. Similarly estimates φu of unconditional performance were obtained from T trials of R replicates each. In the second run hold-out, leave-one-out and leave-v-out crossvalidation based performance was assessed from the original dataset - also averaged over *100* bootstrap replicates. Resubstitution naturally would produce constant results and therefore need not be replicated. The difference between the conditional and unconditional estimates φc and φu respectively and the hold-out, leave-one-out and leave-v-out estimates averaged over *100* replicates produce estimates of conditional and unconditional bias of $\varepsilon_{counting}$, $\varepsilon_{posterior}$, and η .

10.2 Tuning of discriminant procedures

The following describes fine tuning of the kernel density estimation based discriminant procedure and the Hills distance procedure. In the former a choice of the smoothing parameter λ has to be made and for the latter an extension to more than *2* populations has to be derived. For the remaining discriminant procedures for discrete data the application is as outlined in chapters 4 and 6 on direct and indirect discriminant procedures.

---

[31] Further information on this and the *IRIS* data set is given in chapter 11.

10.2.1 Kernel density estimation based procedure

The kernel density estimate used in the comparative analyses is based on a formula for unordered q-variate categorical data $\mathbf{X}=X_1,\ldots,X_q$ (Aitchison and Aitken, 1976) which allows the specification of population specific bandwidth parameters λ_i. A modification that allows the further specification of separate bandwidth parameters λ_{ik} to allow for heteroscedasticity among the variables (Titterington, 1980) - although computationally equally easily achieved - is not adopted in order to achieve sufficient smoothing (see the cautionary note by McLachlan, 1992, p 297). The kernel estimate of the i^{th} population specific density $f_i(\mathbf{x})$ is

$$\hat{f}_i(\mathbf{x}) = \frac{1}{n_i} \sum_{j=1}^{n_i} \prod_{k=1}^{q} \left[\lambda_i^{1-|x_{jk}-x_{lk}|} \left\{ \left(1 - \lambda_i\right) / \left(c_k - 1\right) \right\}^{|x_{jk}-x_{lk}|} \right] \qquad (10.2\text{-}1)$$

where c_k is the number of discrete categories of the k^{th} component of $\mathbf{x}$, n_i is the sample size in population Π_i and λ_i is the population specific smoothing parameter for which

$$\max_k \frac{1}{c_k} \leq \lambda_i \leq 1 \qquad (10.2\text{-}2)$$

gives lower and upper bounds that ensure that the kernel has the desirable property of integrating to unity. In the case of multivariate binary data the above expression reduces to

$$\hat{f}_i(\mathbf{x}) = \frac{1}{n_i} \sum_{j=1}^{n_i} \prod_{k=1}^{q} \left[\lambda_i^{1-|x_{jk}-x_{lk}|} \left\{ 1-\lambda_i \right\}^{|x_{jk}-x_{lk}|} \right]. \qquad (10.2\text{-}3)$$

The smoothing parameters were determined iteratively by inspecting values of λ_i that maximise the performance of

discriminant procedures for a given dataset. Values of $\lambda_i = 0.8$ were adopted throughout. The above settings constitute a considerable compromise not so much because only one type of kernel is used for all datasets but because the smoothing parameter λ_i is fixed throughout. As was pointed out in chapter 5 a greater role is played by the smoothing parameter λ_i than the type of kernel used.

10.2.2 Hills distance procedure

Hills (1967) gives an expression for a distance measure (see chapter 6) for *3* populations as follows

$$\Delta_s^{(3)} = \sum_{j=1}^{s} \left\{ \sum_{i=1}^{3} \frac{(p_{ij} - \frac{1}{2} p_j)^2}{p_j} \right\} ; \quad p_j = \sum_{i=1}^{g} p_{ij} . \qquad (10.2\text{-}4)$$

In order to apply this discriminant procedure also to situations with more than *3* populations expression 10.2-4 had to be generalised.

This is achieved by

$$\Delta_s^{(g)} = \frac{2 \left[\sum_{i=1}^{g} p_{ij} - \sum_{k<l} p_{kj} p_{ij} \right]}{\sum_{i=1}^{g} p_{ij}} ; \quad p_j = \sum_{i=1}^{g} p_{ij} . \qquad (10.2\text{-}5)$$

Henceforth expression 10.2-5 will be used in computing distances for the Hills procedure.

10.2.3 Modification of distributional distance

In chapter 6 it was pointed out that the distributional distance procedure due to Dillon and Goldstein (1978) is not sensitive to the relative weights of state probabilities, p_{ij}, as the differences $d_j^{Matusita} = \left(\sqrt{p_{1j}} - \sqrt{p_{2j}}\right)^2$ may lead to identical values for different levels of p_{ij}. This may result in a disproportionate exaggeration of differences between cells with small frequencies. Thus the distributional distance model in its original form as proposed by Dillon and Goldstein would be expected to overemphasise differences that ultimately are of little consequence for estimates of the misallocation error. This is because the misallocation error is primarily affected by the allocations that a given discriminant procedure specifies for those cells exhibiting higher relative frequencies. To inspect this a slight modification to expression 6.1-3 is carried out as follows

$$d_{modified}^{Matusita} = \sum_{j=1}^{s} \frac{n_{1j}+n_{2j}}{n_1+n_2} \left(\sqrt{p_{1j}} - \sqrt{p_{2j}}\right)^2 . \qquad (10.2\text{-}6)$$

The new factor, $\frac{n_{1j}+n_{2j}}{n_1+n_2}$, in expression 10.2-6 weights the squared differences in cell proportions, $\left(\sqrt{p_{1j}} - \sqrt{p_{2j}}\right)^2$, in proportion to their average relative frequencies. In the following references to the originally proposed distributional distance procedure are indicated by "*DD1*" while the modified version will be called "*DD2*".

10.3 Sampling from discrete populations

When sampling from discrete populations it may happen that the drawn sample $\{x_n\}$ will not exhibit the full range of the sample space. Put differently, the maximum number of

possible discrete states $s = \prod_{k=1}^{q} l_k$ where l_k is the number of levels for the k^{th} of q variables may not be represented by every sample, especially if the sample size is small. In the case of continuous data this is the rule rather than the exception. When dealing with survey data where the number of categories of observed variables is low (frequently no more than *4* levels are encountered) the sample data will normally show observations for every one of the s states. In the bootstrap and other crossvalidation techniques employed therefore the assumption is that the samples for each population consist of at least *1* observation per state. Thus, the artificially generated bootstrap replicates will contain only the states given in the original datasets.

It was further decided to consider *separate sampling* frames instead of *mixture sampling* frames[32]. In medical research interest lies typically in discriminating between patients and controls often stemming from case-control studies. This indicates separate sampling frames where the number of observations per group is fixed in advance and independent estimation of prior probabilities π_i exist. For this reason all examples are analysed with the assumption of separate sampling. This omission is felt to be consistent with the central research theme being selection of optimal procedures for discrete data rather than provision of selection guides for different sampling schemes.

For sampling purposes a binomially distributed random variable $Y_{ij} \sim \mathcal{B}in(n_{ij}, 1/2)$ was chosen thus producing on average equally sized cell frequencies for training and test sets.

---

[32] In mixture sampling objects are selected at random from the population consisting of g groups. Estimates of prior probabilities are thus derived from the sample as $\hat{\pi}_i = n_i / n,\ i=1,\ldots,g.$

Note that for the misallocation error $\varepsilon^{counting}$ defined in chapter 8 the indicator variable $\gamma_i(\mathbf{x}_j)$ in expression 8.2-2 is set to unity only for correct allocations. This means that the error rate $\varepsilon^{counting}$ which is *1 - ccr* will also increase if no allocation is made due to insufficient relative difference in posteriors. $\varepsilon^{counting}$ is thus deliberately constructed to rise with increasing classification threshold τ.

Note also that expression 8.4-1 in chapter 8 for the η-criterion implies that objects not allocated because they do not satisfy the minimum posterior threshold also lead to reducing η. By setting the posteriors $f(\Pi_i|\mathbf{x})$ in expression 8.4-2 to *1* and alternately setting the indicator $\xi_i(\mathbf{x}_j)$ of expression 8.4-1 to either of its extreme values *-1* and *+1* it becomes clear that for each population the theoretical range of values for the raw η' criterion lies within the interval $\{-1,+1\}$. To make η' comparable to the respective ranges obtained by $\varepsilon^{counting}$ and $\varepsilon^{posterior}$ under better-than-chance-allocation a simple transformation is used. The definition

$$\eta = (\eta' + 1) / 2 \qquad (10.4\text{-}1)$$

achieves a mapping of η' onto the interval $\{0,1\}$ which is of the same range as for the error rates. This mapping onto $\{0,1\}$ is not strictly essential as an $\varepsilon^{counting}$ of $\varepsilon^{posterior}$ value of *0* have a different meaning than an η value of *0*. The remapping was done however in order to achieve a general comparison of variances between the error rates and η. Of course it is possible to reduce the variance further by rescaling η onto an even smaller smaller interval.

10.5 Deriving η for indirect procedures

Calculation of performance criteria in the classical direct situation is straightforward. Consider the left hand side of figure 10.5-1. Given distributional information $F(\mathbf{x},\theta)$ and labelled sample data $\{y,\mathbf{x}\}$, where y is the population indicator, the posteriors $\hat{h}_i(\mathbf{x})$ are estimated first. Next allocations are made based directly on these posteriors. Finally performance criteria are derived by counting misallocations for $\varepsilon^{counting}$ and by averaging $\hat{h}_i(\mathbf{x})$ as outlined in chapter 8 for $\varepsilon^{posterior}$ and η. The indirect procedures (right hand side of figure 10.5-1) present a problem because their estimation step provides population distances, object distances, classification trees or neural network weights (synapse strengths). Computation of $\varepsilon^{counting}$ is achieved as for direct procedures but for obtaining suitable and comparable estimates of $\varepsilon^{posterior}$ and η independent posteriors are required. For all indirect procedures it was decided to obtain these from application of the multinomial discriminant procedure as a substitute. They are termed *pseudo-posteriors*.

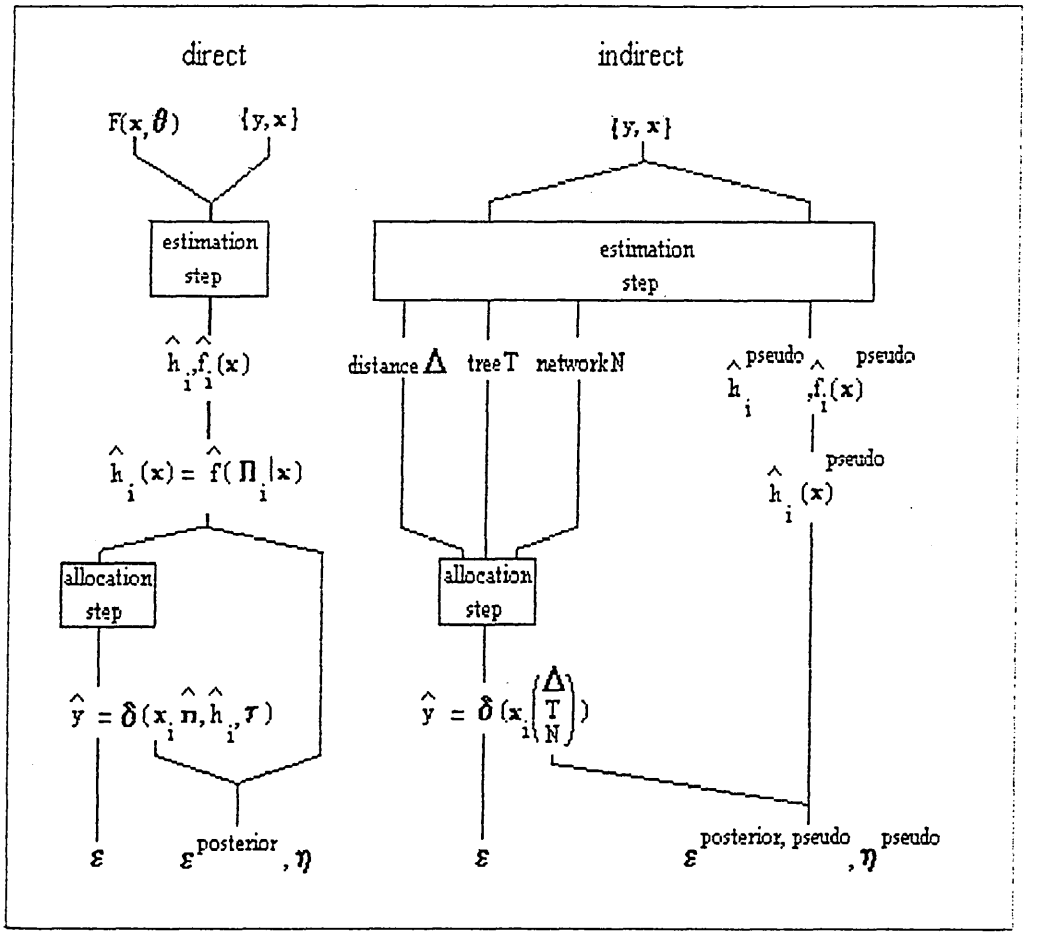


Figure 10.5-1: Performance estimation flowchart

The multinomial procedure was chosen because it is characterised by a high number of parameters, the cell frequencies p_{ij}. Thus, posterior estimates based on the multinomial model allow a fine tuning to differences in empirical distributions. This adjustment also agrees in spirit with the indirect procedures that are generally very computer intensive and tend to provide slightly overspecified solutions. Indirect distance based procedures, such as the centroid procedure or Goldstein and Dillon's distributional distance procedure, generate either individual distances for each object or interpopulation distances that are sensitive to even slight changes in state probabilities p_{ij}. Recursive partitioning procedures are by design such that they produce an overspecified solution unless deliberately restricted by for instance *pruning* of trees as in *CART*. Artificial neural networks

similarly are characterised by a high number of parameters (chapter 6).

10.6 Distribution of relative posterior differences

To analyse the behaviour of a discriminant procedure performance criteria (proportion of cases classified, standard error rates and posterior error rates) are computed as percentages for a range of classification thresholds as outlined in chapter 9. With increasing values of τ the proportion of rejections as well as the error rates will also increase[33].

Several values for τ were tested on a subset of the analyses. The results initially showed that fairly smooth curves could be obtained by varying τ from *0.00* to *1.00* in steps of *0.05* thus giving *21* ordinates for each performance curve. Use of more than *21* ordinates would have resulted in additional computation time. During the phase of actual application (chapter 14) however it became evident that a higher resolution at the upper end above $\tau = 0.90$ was essential in order to discriminate between plots of $f(\tau)$ against τ. In all analyses therefore performance criteria for discriminant procedures are calculated for a standard set of *21* threshold values of τ, and additionally several extra ordinates around $\tau = 0.95$ generally spaced at intervals of $\tau = 0.010$.

The distribution $f(\tau)$ will generally be unknown. It can however be estimated from the given training sample t. As outlined in chapter 9 τ depends on the posteriors $h^{(1)}$ and $h^{(2)}$ via

$$\tau = \frac{h^{(1)} - h^{(2)}}{h^{(1)}} \qquad (10.6\text{-}1)$$

---

[33] The error rates are defined such that they include the rejected cases. Hence, there will be some correlation between rejections and misallocation errors.

As $h_i = f(\Pi_i|\mathbf{x})$ is model dependent it follows that estimates of $\hat{F}(\tau)$ can only be unbiased under correct model assumptions. Assume that such conditions exist, i.e. that $E[\hat{F}(\tau)] = f(\tau)$ such that $\hat{F}(\tau)$ is an unbiased estimator of $f(\tau)$. Then $f(\tau)$ can be estimated using bootstrap techniques. This is achieved by generating the bootstrap distribution $\mathbf{t}^*$ of the training data t. For the k^{th} bootstrap realisation $\mathbf{t}_k^*$ the posteriors are next estimated under the respective data model. The above steps are repeated a sufficient number of times and frequencies for the posteriors averaged.

In the present example *100* bootstrap replications were used for the reasons given in section 10.1. The distribution of τ was discretised into *20* equidistant intervals of length *0.05* plus some additional points (see above) and plotted as a histogram for each procedure applied to a selection of the datasets.

10.7 Program logic

In order to program the respective discriminant procedures and estimation of performance criteria under various crossvalidation options the Fortran language was used (*IBM* professional *FORTRAN* Ryan-McFarland Corporation Compiler Version 1.00 (1984))[34]. The present section outlines the logic of operation of the *DISCRIM* program developed to estimate performance criteria for the research. A complete listing of the source code is available from the author. The main driver program contains 4 different parts (table 10.7-1). Part 1 initialises various parameters and estimates the cumulative distribution function for later bootstrapping. Part 2 estimates the empirical density distribution $f(\tau)$ of relative differences in the two

[34] This Fortran language is designed according to the specifications of the American National Standard Programming Language FORTRAN/77 (ANSI X3.9S - 1978) as understood by IBM. In addition IBM Professional FORTRAN contains many useful extensions to that language.

largest posteriors as outlined in chapter 9. Part 3 contains a major loop that iterates through values of τ from *0.00* to *1.00* in increments of *0.05* or smaller thus providing a smooth metric for the entire range of classification thresholds (see chapter 9). This loop embraces estimation of performance criteria for various crossvalidation schemes (step 3.1) and estimation of expected conditional and unconditional performance for bias and variance calculation (step 3.2). Part 4 outputs all results to files on a disc drive. The program logic in table 10.7-1[35] shows operation of the *DISCRIM* program for a given dataset (real or artificial) and for a given discriminant procedure. Two further loops contained in the source code[36], yet not shown in table 10.7-1 control the iteration of the basic program through all datasets and relevant procedures.

[35] This "table" spreads across several pages.

[36] available from the author

The following four steps of the program logic spanning several pages constitute table 10.7-1.

(1) Compute estimated empirical cumulative distribution function $\hat{F}(\mathbf{x})$ as a basis for nonparametric bootstrap.

(2) Loop_1 through $I = 1$ to N_1

- generate i^{th} bootstrap sample $\mathbf{t}_i^*$ from $\hat{F}(\mathbf{x})$
- estimate posteriors or pseudo posteriors for indirect procedures
- compute distribution of relative differences $f(\tau | \mathbf{t}_i^*)$

End of Loop_1

(3) Loop_2 through $\tau = 0.00$ to 1.00 by increments of 0.05 or smaller

- initialise $\varepsilon^{\text{counting}}$, $\varepsilon^{\text{posterior}}$ and η

(3.1) Estimation of $\varepsilon^{\text{counting}}$, $\varepsilon^{\text{posterior}}$ and η

(3.1.1) Resubstitution crossvalidation

- Estimate allocations from entire original dataset and compute performance criteria by testing the allocation rule δ on same dataset.

(3.1.2) Leave-one-out crossvalidation

Loop_3 through $I = 1$ to number of groups G

Loop_3 through $J = 1$ to number of states S

- Reduce count n_{ij} to $\max(n_{ij}-1, 0)$
- Estimate allocations from reduced sample

End of Loop_3

Average performance criteria over GS cycles.

(3.1.3) Hold-out crossvalidation

Loop_4 through $I = 1$ to N_3

- Split original entire dataset randomly into equally sized training and test sets
- Estimate allocation rule δ from training set and apply it to test data
- Compute performance criteria

End of Loop_4

Average performance criteria over N_j cycles.

(3.1.4) Leave-V-out crossvalidation

- Loop_5 through $I = 1$ to N_3
- Loop_5 through $K = 1$ to N/V
  - Reduce original sample data by V (here this corresponds to 10 percent)
  - Save remaining V observations as test data
  - Estimate allocation rule δ from training data
  - Apply δ to test data
  - Estimate performance criteria

End of Loop_5

Average performance criteria over N_j cycles.

(3.2) Estimation of conditional and unconditional performance for bias and variance calculations

(3.2.1) Conditional performance

Estimate allocation rule δ from original data.

- Loop_6 through $I = 1$ to N_1
  - generate new bootstrap sample t_i^*
  - apply δ to this new sample
  - compute performance criteria

End of Loop_6

Average performance criteria over N_j cycles to obtain expected conditional estimates as well as variance estimates.

(3.2.2) Unconditional performance

Loop_7 through $I = 1$ to N_2

- generate bootstrap sample t_i^*
- estimate bootstrap allocation rule δ_i^*

Loop_8 through $J = 1$ to N_1

- generate further bootstrap sample t_j^*
- apply δ_i^* to t_j^*
- estimate performance criteria

End of Loop_8

Average performance criteria over N_1 cycles

End of Loop_7

Average again over N_2 cycles

End of Loop_2

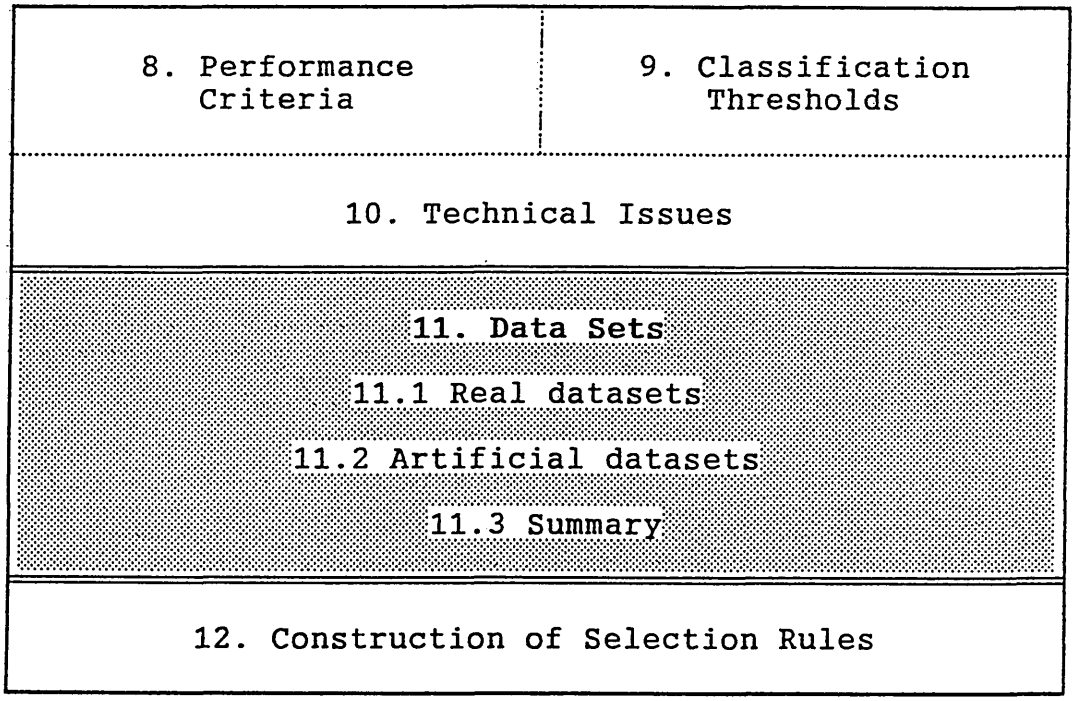
(4) Output results to file for further processing

Table 10.7-1: Program logic

I: INTRODUCTION

II: REVIEW

III: METHOD



IV: RESULTS

V: DISCUSSION

Without loss of generality the following assumptions are made concerning the data. Firstly it will be assumed that all training data consist of correctly labelled multivariate discrete objects thus constituting the *supervised learning* situation as it is known in the pattern recognition terminology. Secondly, datasets will be assumed to consist of a moderate number of predictor variables in the range of *2* to *8*. This further largely obviates the need for special consideration of the *selection of variables*. Thirdly, no missing data are assumed present. This last assumption is perhaps the most stringent in realistic settings. The third restriction was however adhered to for two reasons: (a) with small datasets one may assume in general a greater degree of completeness than with large datasets, and (b) in the case of missing data one of the common approaches is to introduce additional levels for missing values. The resultant discrete dataset may then be treated similarly to one without missing data. The above restrictions were set in order to allow more room for the central issue of quantifying and analysing the distribution of posterior probabilities with a view to constructing suitable performance criteria.

Section 11.1 contains a detailed description of all *real* datasets referred to in chapters 14 and 15. Other datasets, by contrast, are only briefly described. Further information on these real datasets as well as complete enumerations of absolute and relative frequencies for all discrete states are available from the author. It is apparent from the examples given in the following that the majority are of a fairly simple nature exhibiting generally few independent *predictor* variables. Frequently also variables which at first sight might be considered continuous are presented as ordinal or even dichotomous data. This is the case for the *CHD* dataset of Cornfield (1962). This example is typical of those encountered in medical contexts. In the case of the *CHD* data the

continuous variables serum cholesterol and blood pressure were grouped into four categories each corresponding to a prior perceived risk of coronary heart disease.

Section 11.2 gives detail on *artificial* datasets that were generated according to given characteristics in order to inspect possible consequences for procedure selection. These datasets are similarly presented with additional information about how they were generated. Again the emphasis on detail is greatest for the datasets analysed in chapters 14 and 15. Further information on the other datasets is available from the author.

Most of the real datasets are taken from the published literature, some were contributed by the author. The state probabilities given in the following tables for most of the datasets in the columns headed p_1, p_2, ... ,p_g are truncated to *3* decimal digits and may therefore differ slightly from those published in the literature. For the actual calculations, however, the state probabilities were recalculated within an accuracy of *8* decimal digits corresponding to the *REAL*4* data type in *FORTRAN/77* (see also chapter 10 section 7). Simulations were carried out adopting the procedure of Gilbert (1968) and Moore (1973). All descriptions of datasets contain a summary table of basic characteristics within populations (Π_i) and across (*total*). For all datasets analysed in chapters 14 and 15 further diagnostics were carried out including group specific and overall rank order correlations (upper triangular matrices R_i and R) as well as loglinear analysis (see chapter 4) in order to inspect for any significant effects between the predictor variables and the dependent group variable. The correlations are based on Kendall's (1971) nonparametric measure of association, tau_b, ($-1 \leq tau_b \leq 1$). This measure is based on the number of concordant and discordant pairs of observations and uses a correction for tied pairs. Correlations significant at the *0.05* level are printed in bold type.

Table 11.1-1 shows abbreviated names of real datasets used in comparative analyses classified by number of populations (down) and highest level of measurement attained by any one variable (across). The code numbers indicate the number of variables involved at each respective measurement level. The three digits respectively give the number of binary, nominal and ordinal variables in each dataset. The *KRETSCHM* dataset for instance contains *3* dichotomous and *1* nominal variable recorded for discrimination between two groups.

	dichotomous	nominal	ordinal
2 pops	200 *LIZARD* 200 *SEEDLING* 200 *VIRGIN* ..... 400 *BREAST* 400 *CESAR4*	020 *VOTING*	002 *CHD* 002 *ESTEEM* ..... 301 *KRETSCHM* 202 *COLLEGE* ..... 206 *CREDIT*
3 pops	..... 400 *GRADE*		004 *IRIS* .....
4 pops			002 *EDUC*

Table 11.1-1: Summary of real datasets

The following list gives the key characteristics of the real datasets used in alphabetical order. Descriptions follow an identical pattern. In all cases a subset of actual values is included for illustration.

11.1.1 *BREAST* data

The dichotomous response, survival after breast cancer, is predicted from the *4* dichotomous predictors: X_1 diagnostic centre (Tokyo or Glamorgan), X_2 age (under *50* years or over *70* years), X_3 inflammation (minimal or greater) and X_4 appearance (benign or malignant). The data are reported in Bishop, Fienberg and Holland (1975).

		Π_1	Π_2	total
size	n_i	*554*	*210*	*764*
	π_i	*.725*	*.275*	
mean	*1*	*1.861*	*2.038*	*1.910*
	2	*1.769*	*1.910*	*1.808*
	3	*1.199*	*1.210*	*1.202*
	4	*1.583*	*1.462*	*1.550*
std	*1*	*.821*	*.782*	*.813*
	2	*.694*	*.743*	*.710*
	3	*.399*	*.408*	*.401*
	4	*.494*	*.500*	*.498*

Table 11.1-2: Characteristics of *BREAST* data

11.1.2 *CESAR4* data

The research aim is the prediction of delivery by caesarean section from *4* dichotomous predictors. The data (restricted to live singleton births) are extracted from the perinatal survey of Lower Saxony in the years 1986 to 1988.

response — delivery by caesarean section
- (*1*) caesarean section
- (*2*) vaginal delivery

predictors — dichotomous
- (X_1) position of fetus in the womb
  - (*0*) vertex
  - (*1*) breech
- (X_2) twin pregnancy
  - (*0*) no
  - (*1*) yes
- (X_3) previous caesarean or uterus surgery
  - (*0*) no
  - (*1*) yes
- (X_4) presence of placental insufficiency
  - (*0*) no
  - (*1*) yes

state	pattern	n1	n2	p1	p2
1	0000	78	1181	.345	.896
2	0001	13	25	.057	.018
3	0010	14	18	.061	.013
.	.	.	.	.	.
.	.	.	.	.	.
12	1110	2	0	.008	.000

		Π_1	Π_2	total
size	n_i	226	1318	1544
	π_i	.146	.854	
mean	1	.447	.069	.124
	2	.133	.002	.021
	3	.137	.018	.035
	4	.097	.021	.032
std	1	.498	.254	.330
	2	.340	.048	.145
	3	.345	.131	.184
	4	.297	.142	.175

$$R_1 = \begin{pmatrix} 1.000 & -.063 & -.100 & -.085 \\ & 1.000 & .147 & -.084 \\ & & 1.000 & -.044 \\ & & & 1.000 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1.000 & .050 & .032 & -.018 \\ & 1.000 & .115 & -.007 \\ & & 1.000 & .022 \\ & & & 1.000 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.000 & .107 & .067 & .021 \\ & 1.000 & .191 & -.001 \\ & & 1.000 & .026 \\ & & & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2, X_3, X_4
interactions	-

Table 11.1-3: Characteristics of *CESAR4* data

11.1.3 *CHD* data

The research aim is the prediction of imminent coronary heart disease based on ordinal serum cholesterol at four levels and blood pressure also at four levels. The data were compiled by Cornfield (1962) in the context of the Framingham longitudinal study on coronary heart disease (see Dawber, Kannel and Lyell (1963) for details). The data are reported in Fienberg (1980). The most prominent features of the data are a very low prior probability for the *CHD* population yet coupled with a strong interest in

achieving high sensitivity in detection at a low false positive rate (or high specificity).

response — dichotomous
- (*1*) *CHD* present
- (*2*) *CHD* absent

predictors — *2* ordinal
- (X_1) serum cholesterol
  - (*1*) < *200* mg / *100* cc
  - (*2*) *200* - *219* mg / *100* cc
  - (*3*) *220* - *259* mg / *100* cc
  - (*4*) ≥ *260* mg / *100* cc
- (X_2) systolic blood pressure
  - (*1*) < *127* mm Hg
  - (*2*) *127* - *146* mm Hg
  - (*3*) *147* - *166* mm Hg
  - (*4*) ≥ *167* mm Hg

state	pattern	n1	n2	p1	p2
1	*11*	*2*	*117*	*.021*	*.094*
2	*12*	*3*	*121*	*.032*	*.097*
.	.	.	.	.	.
16	*44*	*11*	*33*	*.119*	*.026*

		Π_1	Π_2	total
size	n_i	*92*	*1237*	*1329*
	π_i	*.069*	*.931*	
mean	*1*	*3.098*	*2.503*	*2.544*
	2	*2.522*	*2.042*	*2.075*
std	*1*	*1.028*	*1.069*	*1.077*
	2	*1.104*	*.927*	*.948*

$$R_1 = \begin{pmatrix} 1.000 & .042 \\ & 1.000 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1.000 & .087 \\ & 1.000 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.000 & .095 \\ & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2
interactions	-

Table 11.1-4: Characteristics of *CHD* data

11.1.4 *COLLEGE* data

In a study of randomly selected cohort of Wisconsin high school seniors, Sewell and Shah (1968) explored the relationship among five variables: X_1 sex (male, female), X_2 intelligence (high, upper middle, lower middle, low) as measured by the Hemmon-Nelson Test of Mental Ability, X_3 parental encouragement (low, high), X_4 socioeconomic status (high, upper middle, lower middle, low) and college plans (yes, no). In the current application to discriminant analysis interest is in whether the former four variables allow prediction of the dichotomous response, intent to go to college. The data are given in Fienberg (1980).

		Π_1	Π_2	total
size	n_i	*3376*	*6942*	*10318*
	π_i	*.327*	*.673*	
mean	*1*	*1.455*	*1.546*	*1.516*
	2	*3.092*	*2.195*	*2.488*
	3	*1.908*	*1.330*	*1.519*
	4	*3.133*	*2.206*	*2.510*
std	*1*	*.498*	*.498*	*.500*
	2	*.973*	*1.053*	*1.110*
	3	*.290*	*.470*	*.500*
	4	*.978*	*1.036*	*1.107*

Table 11.1-5: Characteristics of *COLLEGE* data

11.1.5 *CREDIT* data

The research aim is prediction of credit worthiness of bank customers based on demographic data and account running behaviour, dichotomous response: credit worthiness, Fahrmeir et al (1984). Every bank will be faced with the problem of allocating a potential new customer to either the no-hassle straightforward paying group or to the problem group. The original data are taken from the database of a large south German bank. They have been simplified for the purposes of the present study by reducing the number of predictors to *6* from originally *20*. This was done by considering ranked univariate χ^2

statistics. Further the number of levels of these 6 selected predictors was reduced to three throughout to keep the number of discrete states to a manageable quantity. The data were kindly supplied by L. Fahrmeir.

response — credit worthiness
- (1) not credit worthy
- (2) credit worthy

predictors
- (X_1) currently held account (ordinal)
  - (1) more than 200 DM in credit over 1 year
  - (2) no current account with this bank
  - (3) up to 200 DM in credit
- (X_2) past paying morale (ordinal)
  - (1) history of past irregular payments
  - (2) current credits running satisfactorily
  - (3) past credits settled satisfactorily
- (X_3) savings (dichotomous)
  - (1) none or below 500 DM
  - (2) 500 DM and more
- (X_4) purpose of credit (nominal)
  - (1) educational or essential household goods
  - (2) luxury items
  - (3) repair or other
- (X_5) assets (ordinal)
  - (1) none or at most a car
  - (2) building society or life insurance policy
  - (3) house or property owner
- (X_6) employment (ordinal)
  - (1) unemployed or employ for at most one year
  - (2) in employment between 1 and 7 years
  - (3) employed beyond 7 years

state	pattern	n1	n2	p1	p2
1	111113	0	2	.000	.002
2	111211	7	0	.023	.000
3	111212	1	0	.003	.000

.	.	.	.	.	.
.	.	.	.	.	.
236	233333	0	5	.000	.007

		Π_1	Π_2	total
size	n_i	300	700	1000
	π_i	.300	.700	
mean	1	1.200	1.567	1.457
	2	1.977	2.276	2.186
	3	.990	2.296	2.204
	4	1.897	2.199	2.108
	5	2.357	2.573	2.508
	6	1.440	1.799	1.691
std	1	.401	.496	.498
	2	.729	.676	.706
	3	.587	.558	.584
	4	.793	.809	.816
	5	.867	.748	.792
	6	.758	.929	.896

$$R_1 = \begin{pmatrix} 1.000 & .048 & .118 & .004 & .034 & .156 \\ & 1.000 & -.043 & -.112 & .199 & .047 \\ & & 1.000 & -.045 & .028 & .000 \\ & & & 1.000 & -.009 & -.009 \\ & & & & 1.000 & -.056 \\ & & & & & 1.000 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1.000 & -.002 & .123 & .070 & .028 & .137 \\ & 1.000 & .068 & -.100 & .193 & -.099 \\ & & 1.000 & -.044 & .005 & -.041 \\ & & & 1.000 & -.017 & .017 \\ & & & & 1.000 & -.065 \\ & & & & & 1.000 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.000 & .070 & .187 & .100 & .067 & .190 \\ & 1.000 & .047 & -.067 & .205 & -.026 \\ & & 1.000 & -.007 & .038 & .009 \\ & & & 1.000 & .006 & .037 \\ & & & & 1.000 & -.040 \\ & & & & & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2, X_3, X_4
interactions	-

Table 11.1-6: Characteristics of *CREDIT* data

11.1.6 *EDUC* data

The research aim is the prediction of occupational group later attained in life on the basis of information gathered during primary education. *2* ordinal variables X_1 educational level and X_2 aptitude level (as measured by the scholastic aptitude test by Beaton (1975)) are used for prediction. The survey was conducted by National Bureau of Economic Research, Thorndike and Hagen (1959). The data are given in Fienberg (1980).

response	occupational group (ordinal)		
(*1*)	self-employed, business		
(*2*)	self-employed, professional		
(*3*)	teacher		
(*4*)	salary-employed		
predictors	*2* ordinal		
(X_1)	educational level		
	(*1*)	E*1*	lowest
	(*2*)	E*2*	low
	(*3*)	E*3*	high
	(*4*)	E*4*	highest
(X_2)	aptitude level		
	(*1*)	A*1*	lowest
	(*2*)	A*2*	low
	(*3*)	A*3*	medium
	(*4*)	A*4*	high
	(*5*)	A*5*	highest

state	pattern	n1	n2	n3	n4	p1	p2	p3	p4
1	*11*	*42*	*1*	*0*	*172*	*.050*	*.003*	*.000*	*.057*
2	*12*	*55*	*2*	*0*	*151*	*.065*	*.006*	*.000*	*.050*
3	*13*	*22*	*8*	*1*	*107*	*.026*	*.027*	*.004*	*.035*
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
20	*54*	*5*	*19*	*14*	*79*	*.005*	*.066*	*.059*	*.026*

		Π_1	Π_2	Π_3	Π_4	total
size	n_i	834	286	235	2998	4353
	π_i	.192	.066	.054	.689	
mean	1	2.712	3.035	2.838	2.784	2.789
	2	2.120	3.615	3.877	2.377	2.490
std	1	1.082	1.114	1.008	1.148	1.128
	2	.898	.664	.441	1.048	1.083

$$R_1=\begin{pmatrix} 1.000 & .125 \\ & 1.000 \end{pmatrix}, \quad R_2=\begin{pmatrix} 1.000 & .043 \\ & 1.000 \end{pmatrix}$$

$$R_3=\begin{pmatrix} 1.000 & .013 \\ & 1.000 \end{pmatrix}, \quad R_4=\begin{pmatrix} 1.000 & .166 \\ & 1.000 \end{pmatrix}$$

$$R=\begin{pmatrix} 1.000 & .152 \\ & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2
interactions	-

Table 11.1-7: Characteristics of *EDUC* data

11.1.7 *ESTEEM* data

Rosenberg (1962) reports results of a survey assessing respondent's self-esteem as either (*1*) high or (*2*) low. The current aim is prediction of self-esteem from X_1 religion (catholic, jewish and protestant) and father's education (8th grade or less, some high school education, high school graduate, some college education, college graduate, postgraduate education). The data are given in Fienberg (1980).

		Π_1	Π_2	total
size	n_i	*2614*	*1136*	*3750*
	π_i	*.697*	*.303*	
mean	*1*	*1.922*	*1.979*	*1.940*
	2	*3.060*	*2.905*	*3.013*
std	*1*	*.910*	*.943*	*.920*
	2	*1.488*	*1.409*	*1.466*

Table 11.1-8: Characteristics of *ESTEEM* data

11.1.8 *GRADE* data

The research aim is prediction of performance on a test in terms of class of school grade on the basis of demographic data. The data are reported by Goldstein & Dillon (1978).

response school grade class (ordinal)
- (*1*) low
- (*2*) medium
- (*3*) high

predictors
- (X_1) sex
  - (*0*) male
  - (*1*) female
- (X_2) intelligence quotient
  - (*0*) lower than *100*
  - (*1*) greater than *100*
- (X_3) social class
  - (*0*) lower
  - (*1*) higher
- (X_4) family size
  - (*0*) up to *2* children
  - (*1*) *3* or more children

state	pattern	n1	n2	n3	p1	p2	p3
1	*0000*	*1*	*19*	*2*	*.083*	*.253*	*.153*
2	*0001*	*0*	*6*	*1*	*.000*	*.080*	*.076*

3	0010	2	12	3	.166 .160 .230
.	.	.	.	.	. . .
.	.	.	.	.	. . .
13	1111	1	1	0	.083 .013 .000

		Π_1	Π_2	Π_3	total
size	n_i	12	75	13	100
	π_i	.120	.750	.130	
mean	1	.417	.467	.549	.470
	2	.333	.053	.000	.080
	3	.750	.427	.231	.440
	4	.750	.200	.077	.250
std	1	.515	.502	.519	.502
	2	.492	.226	.000	.273
	3	.452	.498	.439	.499
	4	.452	.403	.277	.435

$$R_1=\begin{pmatrix} 1.000 & .478 & -.293 & .488 \\ & 1.000 & -.408 & .408 \\ & & 1.000 & .111 \\ & & & 1.000 \end{pmatrix} \quad R_2=\begin{pmatrix} 1.000 & .016 & .220 & .067 \\ & 1.000 & -.085 & .030 \\ & & 1.000 & -.094 \\ & & & 1.000 \end{pmatrix}$$

$$R_3=\begin{pmatrix} 1.000 & -.592 & -.312 & - \\ & 1.000 & - & - \\ & & 1.000 & .158 \\ & & & 1.000 \end{pmatrix} \quad R=\begin{pmatrix} 1.000 & .092 & .053 & .058 \\ & 1.000 & -.039 & .255 \\ & & 1.000 & .047 \\ & & & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_2
interactions	-

Table 11.1-9: Characteristics of *GRADE* data

11.1.9 *IRIS* data

The research aim is the prediction of type of type of iris flower from sepal length and width as well as petal length and width. The original data are due to Fisher (1936). For the present purposes the continuous variables were discretised into *3* levels depending on the respective *25*th and *75*th percentiles[37]. This was done in order to inspect

[37] the zero standard deviations for X_3 and X_4 in population

for robustness of the linear discriminant function under departures from normality.

response species (nominal)
(*1*) versicolor
(*2*) setosa
(*3*) virginica
predictors
(X_1) sepal length
(*1*) < *25*th percentile
(*2*) interquartile range
(*3*) > *75*th percentile
(X_2) sepal width
(*1*) < *25*th percentile
(*2*) interquartile range
(*3*) > *75*th percentile
(X_3) petal length
(*1*) < *25*th percentile
(*2*) interquartile range
(*3*) > *75*th percentile
(X_4) petal width
(*1*) < *25*th percentile
(*2*) interquartile range
(*3*) > *75*th percentile

state	pattern	n1	n2	n3	p1	p2	p3
1	*1111*	*1*	*2*	*0*	*.020*	*.040*	*.000*
2	*1112*	*0*	*1*	*0*	*.000*	*.020*	*.000*
3	*1121*	*0*	*1*	*0*	*.000*	*.020*	*.000*
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
23	*3333*	*0*	*0*	*5*	*.000*	*.000*	*.100*

Π_1 as well as for X_4 in population Π_3 are due to original data values yielding identical discrete codings as a consequence of discretisation with corresponding consequences for the correlation matrices R_1 and R_2

		Π_1	Π_2	Π_3	total
size	n_i	50	50	50	150
	π_i	.333	.333	.333	
mean	1	2.600	1.600	1.880	2.027
	2	1.100	2.200	2.760	2.020
	3	1.000	2.160	2.980	2.047
	4	1.000	2.620	3.000	2.207
std	1	.535	.535	.558	.685
	2	.303	.639	.476	.847
	3	.000	.650	.141	.900
	4	.000	.602	.000	.936

$$R_1=\begin{pmatrix} 1.000 & .257 & - & - \\ & 1.000 & - & - \\ & & 1.000 & - \\ & & & 1.000 \end{pmatrix} \quad R_2=\begin{pmatrix} 1.000 & .350 & .444 & .610 \\ & 1.000 & .377 & .563 \\ & & 1.000 & .543 \\ & & & 1.000 \end{pmatrix}$$

$$R_3=\begin{pmatrix} 1.000 & .408 & .228 & - \\ & 1.000 & .259 & - \\ & & 1.000 & - \\ & & & 1.000 \end{pmatrix} \quad R=\begin{pmatrix} 1.000 & -.218 & -.287 & -.347 \\ & 1.000 & .763 & .802 \\ & & 1.000 & .860 \\ & & & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2
interactions	$X_1 * X_2$

Table 11.1-10: Characteristics of *IRIS* data

11.1.10 *KRETSCHM* data

The research aim is the prediction of a patient's prognosis based on hospital admission data. The dichotomous response is defined as (*1*) alive *40* days after discharge or (*2*) dead. The predictors are X_1 admission status (primary admission, secondary admission, repeated admission), X_2 malfunction of vegetative functions (no, yes), X_3 stiffness (no, yes) and X_4 bed ridden (no, yes). The data are reported by Victor (1976).

		Π_1	Π_2	total
size	n_i	*20*	*20*	*40*
	π_i	*.500*	*.500*	
mean	*1*	*1.650*	*1.400*	*1.525*
	2	*1.200*	*1.600*	*1.400*
	3	*1.050*	*1.350*	*1.200*
	4	*1.250*	*1.500*	*1.375*
std	*1*	*.813*	*.821*	*.816*
	2	*.410*	*.503*	*.496*
	3	*.224*	*.489*	*.405*
	4	*.444*	*.513*	*.490*

Table 11.1-11: Characteristics of *KRETSCHM* data

11.1.11 *LIZARD* data

The research aim is the prediction of species of adult male lizard (either *Sagrei* lizards or *angusticeps* lizards) from *2* dichotomised predictors (X_1 perch diameter less than or greater than *2.5* inches and X_2 perch height less than or greater than *5.0* feet). The data were gathered by Schoener (1968) and are given in Fienberg (1980).

		Π_1	Π_2	total
size	n_i	*165*	*27*	*192*
	π_i	*.859*	*.141*	
mean	*1*	*.800*	*.185*	*.714*
	2	*.618*	*.111*	*.547*
std	*1*	*.401*	*.396*	*.453*
	2	*.487*	*.320*	*.499*

Table 11.1-12: Characteristics of *LIZARD* data

11.1.12 *SEEDLING* data

Wakely (1954) investigated the effect of planting longleaf and slash pine seedlings *1/2* inch too high or too deep in winter upon their mortality in the following autumn. Seedling type X_2 is coded as *0* and *1* for longleaf and

slash, respectively. Depth of planting X_1 is coded as *0* and *1* for too high and too low respectively. The data are reported by Fienberg (1980).

		Π_1	Π_2	total
size	n_i	*69*	*331*	*400*
	π_i	*.173*	*.828*	
mean	*1*	*.232*	*.556*	*.500*
	2	*.246*	*.553*	*.500*
std	*1*	*.425*	*.498*	*.501*
	2	*.434*	*.498*	*.501*

Table 11.1-13: Characteristics of *SEEDLING* data

11.1.13 *VIRGIN* data

In a retrospective study of premarital contraceptive usage, Reiss, Banwart and Foreman (1975) took samples of undergraduate female university students. One sample consisted of individuals who had attended the university contraceptive clinic, and the other was a control group consisting of females who had not done so. The individuals in the two samples were then crossclassified according to their virginity and to their belief of whether extramarital coitus is always or not always wrong. In the present application use of clinic is predicted on the basis of the two dichotomous variables X_1 virginity (yes, no) and X_2 attitude on extramarital coitus (always wrong, not always wrong). The data are given in Fienberg (1980).

		Π_1	Π_2	total
size	n_i	*291*	*123*	*414*
	π_i	*.703*	*.297*	
mean	*1*	*.485*	*.667*	*.539*
	2	*.821*	*.268*	*.657*
std	*1*	*.501*	*.473*	*.499*
	2	*.384*	*.445*	*.475*

Table 11.1-14: Characteristics of *VIRGIN* data

11.1.14 *VOTING* data

Analysis of the effects of area of residence (rural, metropolitan or city) and political affiliation (republican, democrat or other/none) on one's attitude towards a local communal issue (for or against). The data relate to research on voting behaviour conducted by the University of Michigan and reported by Kish (1957).

		Π_1	Π_2	total
size	n_i	*1003*	*1439*	*2442*
	π_i	*.411*	*.589*	
mean	*1*	*1.229*	*2.318*	*1.871*
	2	*1.707*	*2.506*	*2.178*
std	*1*	*.485*	*.788*	*.866*
	2	*.675*	*.618*	*.753*

Table 11.1-15: Characteristics of *VOTING* data

11.2 Artificial datasets

Table 11.2-1 summarises all artificial datasets used in comparative analyses classified by highest level of measurement attained by any one variable. The code numbers indicate the number of variables involved at each respective measurement level. All datasets beginning with *NORMAL* are realisations of simulated data. The sequence *NORMAL11 ... NORMAL17* crosses classification boundaries due to different degrees of *discreteness.*

dichotomous	nominal	ordinal
400 *MA435300*		002 *BANANA*
400 *MA435301*		002 *INTERAC1*
400 *MA435302*		002 *POISSON*
400 *MA435303*		
400 *MA435304*		002 *NORMAL01*
400 *MA435305*		002 *NORMAL02*
400 *MA435306*		002 *NORMAL03*
400 *MA435307*		
400 *MA435308*		002 *NORMAL11*
400 *MA435309*		002 *NORMAL12*
		002 *NORMAL*13
200 *DILLON*		002 *NORMAL*14
	020 *NORMAL*15	
	020 *NORMAL*16	
200 *NORMAL*17		

Table 11.2-1: Artificial data by predictor class

In addition to real datasets artificially generated datasets were used to inspect particular aspects of a discriminant's performance, especially the modelling of interactions. These *artificial* datasets are generated using two methods. The first is achieved by using the Lazarsfeld - Bahadur reparametrisation for joint binomial distributions. This method is used for the *MA435300* .. *MA435309* series.

11.2.1 *BANANA* data

Artificial data constructed such that one population's objects are concentrated in one spot while the other population's objects wrap around (in the shape of a banana) such that a linear separation line cannot be placed without leading to considerable overlap and hence misallocation error. The *2* independent variables X_1 and X_2 are ordinally coded taking on discrete values between *1* and *6*.

		Π_1	Π_2	total
size	n_i	1232	2464	3696
	π_i	.333	.667	
mean	1	2.122	3.812	2.967
	2	2.506	3.606	3.056
std	1	1.347	.842	1.405
	2	1.433	.865	1.305

Table 11.2-2: Characteristics of *BANANA* data

11.2.2 *DILLON* data

Hypothetical data given by Dillon and Goldstein (1978) to demonstrate the distributional distance model. The data are also discussed in chapter 6.

		Π_1	Π_2	total
size	n_i	25	274	299
	π_i	.084	.916	
mean	1	.160	.442	.418
	2	.720	.409	.435
	3	.320	.661	.632
std	1	.374	.498	.494
	2	.458	.493	.497
	3	.476	.474	.483

Table 11.2-3: Characteristics of *DILLON* data

11.2.3 *INTERAC1* data

The research aim is the inspection of performance of discriminant procedures when faced with data containing interactions. The data are constructed such that both population's objects are concentrated in two centroids diagonally placed at opposite corners of a square[38]. The data were already introduced in chapter 2 to illustrate a situation where effective separation can not be achieved

[38] the fact that the correlations within populations are large and significant while the correlation in the pooled sample is negligible is a direct consequence of the design of this dataset

using a single separation line. In the case of the *INTERAC1* dataset separation between both populations can be improved substantially by use of two curvilinear separation lines. The dichotomous response is predicted from X_1 with *4* discrete levels and X_2 with *5* discrete levels. The dataset is shown again in figure 11.2-1.

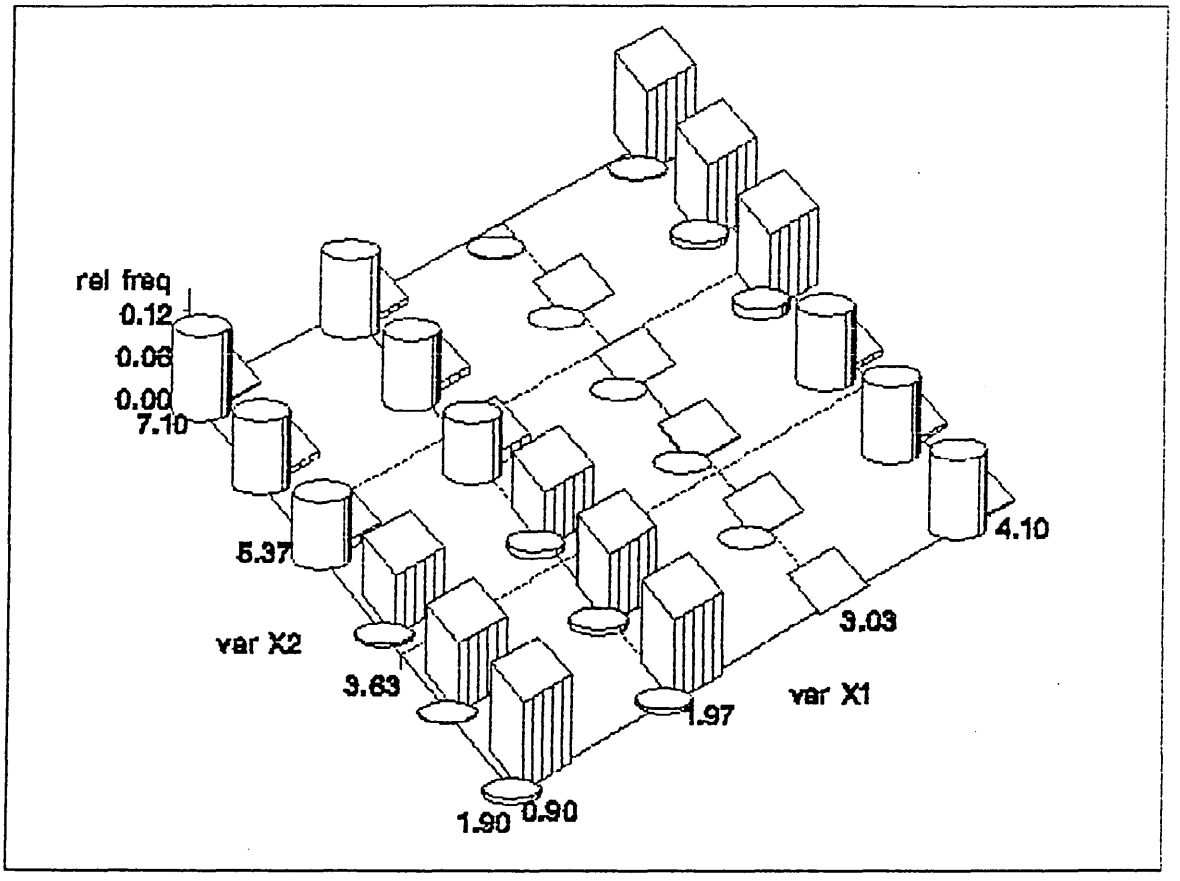


Figure 11.2-1: The artificial *INTERAC1* data

state	pattern	n1	n2	p1	p2
1	*12*	*49*	*1007*	*.005*	*.119*
2	*13*	*17*	*789*	*.002*	*.093*
3	*14*	*27*	*688*	*.003*	*.081*
.	.	.	.	.	.
.	.	.	.	.	.
24	*47*	*19*	*989*	*.002*	*.117*

		Π_1	Π_2	total
size	n_i	8231	8405	16636
	π_i	.495	.505	
mean	1	2.377	2.391	2.384
	2	4.971	4.120	4.541
std	1	1.261	1.258	1.259
	2	1.680	1.718	1.751

$$R_1 = \begin{pmatrix} 1.000 & -.561 \\ & 1.000 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1.000 & .566 \\ & 1.000 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.000 & .010 \\ & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2
interactions	$X_1 * X_2$

Table 11.2-4: Characteristics of *INTERAC1* data

11.2.4 *MA435300 .. MA435309* data

As described in chapter 4 a joint binomial distribution can be expressed in terms of means and correlations between *2* or more variables (Bahadur, 1961a). It therefore lends itself suitably to the task of modelling a variety of distributions with different degrees of complexity. Here the research aim will be in particular the inspection of the Bahadur based discriminant procedures. In the case of *3* and more populations Lancaster's definition has to be used (chapter 4 section 1). The datasets are then generated as follows:

(1) specify the characteristics of the populations in terms of means and correlations

(2) insert these as parameters into the functional definition of the multivariate binomial density

(3) for a given cell - corresponding to one of all possible combinations of independent variables - calculate the density defined by (2)

(4) interpret the resultant vector of density estimates as parameter values of a multinomial distribution and sample from this using a

standard random number generator giving uniform deviates in the interval $\{0,1\}$

(5) repeat step (4) n times where n is the required sample size.

The characteristics (1) were specified in populations Π_1 and Π_2 as follows:

marginal mean levels of individual variables

(Π_1) *0.4, 0.4, 0.4, 0.4*

(Π_2) *0.6, 0.6, 0.6, 0.6*

correlations between two variables at a time

(Π_1) *0.3, 0.3, 0.3, 0.3, 0.3, 0.3*

(Π_2) *0.6, 0.6, 0.6, 0.6, 0.6, 0.6*

correlations between three variables at a time

(Π_1) *0.2, 0.2, 0.2, 0.2*

(Π_2) *0.5, 0.5, 0.5, 0.5*

dataset	means	correlations
MA435300	$\mu_1=.4 \quad \mu_2=.6$	$\rho_1^{(2)}=.3 \quad \rho_2^{(2)}=.6$ $\rho_1^{(3)}=.2 \quad \rho_2^{(3)}=.5$
.	.	. .
.	.	. .
MA435309	$\mu_1=.4 \quad \mu_2=.6$	$\rho_1^{(2)}=.3 \quad \rho_2^{(2)}=.6$ $\rho_1^{(3)}=.2 \quad \rho_2^{(3)}=.5$

Table 11.2-5: Artificial Bahadur data

Table 11.2-5 shows parameters for the series of *10* datasets generated using Monte-Carlo techniques under the *Bahadur-Lazarsfeld* reparametrisation of joint multivariate binary densities[39]. Sample sizes are constant (n = *100*) for

[39] This is defined in terms of mean vectors (μ_i) and correlations between k variables ($\rho_i^{(k)}$). Individual

all datasets with equal priors. An example of such a dataset including codings of variables and key characteristics are given below.

state	pattern	n1	n2	p1	p2
1	*0000*	*16*	*7*	*.160*	*.070*
2	*0001*	*6*	*7*	*.060*	*.070*
3	*0010*	*11*	*15*	*.110*	*.150*
.	.	.	.	.	.
.	.	.	.	.	.
16	*1111*	*14*	*45*	*.140*	*.449*

average over *10* samples		Π_1	Π_2	total
size	n_i	*100*	*100*	*200*
	π_i	*.500*	*.500*	
mean	*1*	*.419*	*.584*	*.502*
	2	*.402*	*.611*	*.507*
	3	*.401*	*.620*	*.511*
	4	*.407*	*.617*	*.512*
std	*1*	*.494*	*.494*	*.501*
	2	*.491*	*.488*	*.501*
	3	*.488*	*.487*	*.454*
	4	*.445*	*.485*	*.499*

$$R_1=\begin{pmatrix} 1.000 & .271 & .298 & .306 \\ & 1.000 & .328 & .325 \\ & & 1.000 & .295 \\ & & & 1.000 \end{pmatrix}$$

$$R_2=\begin{pmatrix} 1.000 & .516 & .510 & .509 \\ & 1.000 & .456 & .464 \\ & & 1.000 & .412 \\ & & & 1.000 \end{pmatrix}$$

$$R=\begin{pmatrix} 1.000 & .413 & .424 & .428 \\ & 1.000 & .421 & .429 \\ & & 1.000 & .384 \\ & & & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2, X_3, X_4
interactions	-

Table 11.2-6: Characteristics of *MA435300* data

The second method of generating artificial data consists of *discretising* a sample of the multivariate normal distribution. This approach is used for the *NORMAL11* .. *NORMAL17* series and for the *NORMAL01*, *NORMAL02* and *NORMAL03* datasets.

components of these vectors are set to identical values, e.g: $\mu_1'=(\mu_{11},\ldots,\mu_{1r})=(0.4,\ldots,0.4)$.

11.2.5 *NORMAL11* .. *NORMAL17* data

As seen in chapter 4 the linear discriminant function based procedure shows considerable robustness to departures from the normality assumptions. In order to inspect this procedure for its performance under extreme conditions it was decided to subject a sequence of specially designed artificial datasets to it. These were constructed such that they ranged from *close to normal* through to *extremely non-normal* by systematically varying the *degree of discreteness* of an originally continuous and normal dataset. To achieve this initially continuous samples of *200* observations, *100* for each of *2* groups with equal priors $\pi_1=\pi_2$ and parameters $\mu_1=\begin{pmatrix}50\\30\end{pmatrix}$ and $\mu_2=\begin{pmatrix}50\\30\end{pmatrix}$ and common covariance matrix $\begin{pmatrix}100 & 0\\0 & 25\end{pmatrix}$ were generated. Subsequently the respective ranges of the independent variates X_1 and X_2 were divided into at most *15* equidistant intervals per variable corresponding to the dataset labelled (*NORMAL11*) and at least *2* intervals corresponding to the dataset labelled (*NORMAL17*) thus yielding *discrete* datasets with the number of discrete states *s* ranging from *96* to *4* respectively. Figure 11.2-2 shows a *3*-dimensional plot of the bivariate *6*-level *discretised NORMAL14* simulated data with relative frequency plotted vertically.

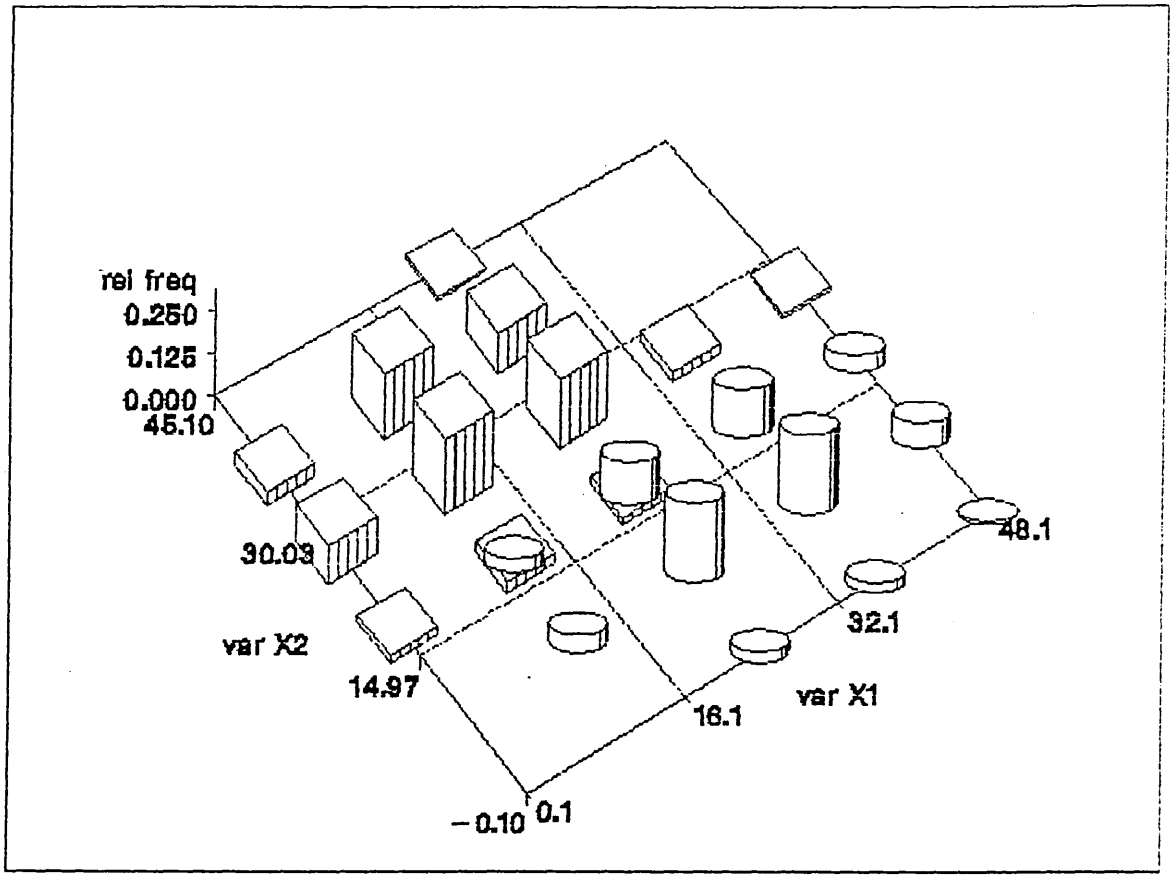


Figure 11.2-2: "Discretised" artificial density

Figure 11.2-2 shows "discretised" bivariate density generated by separate sampling from continuous independent bivariate normal density with means $\mu_1'=(50,30)$ and $\mu_2'=(20,40)$ and common covariance vector $\sigma^{2\prime}=(100,25)$ and subsequently transforming to 6-state "discrete" data with equidistant metric. This yields $X_1 \in \{00, 12, 24, 36, 48\}$ and $X_2 \in \{00, 09, 18, \ldots, 45\}$.

state	pattern	n1	n2	p1	p2
1	0018	0	2	0.000	0.020
2	0027	0	11	0.000	0.110
3	0036	0	4	0.000	0.040
.	.	.	.	.	.
.	.	.	.	.	.
21	4827	0	1	0.000	0.010

		Π_1	Π_2	total
size	n_i	100	100	200
	π_i	.500	.500	
mean	1	30.840	15.240	23.040
	2	11.520	29.700	20.610
std	1	9.983	9.776	12.581
	2	5.287	5.502	10.583

$$R_1=\begin{pmatrix} 1.000 & -.007 \\ & 1.000 \end{pmatrix} \quad R_2=\begin{pmatrix} 1.000 & .0291 \\ & 1.000 \end{pmatrix} \quad R=\begin{pmatrix} 1.000 & -.454 \\ & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2
interactions	-

Table 11.2-7: Characteristics of *NORMAL14* data

11.2.6 *NORMAL01* data

The research aim for this and also for the *NORMAL02* and the *NORMAL03* dataset is the inspection of the behaviour of discriminant procedures under departures from normality. This dataset was generated by discretising a tri-variate normal with original means $\mu_1'=(25,15,45)$ & $\mu_2'=(30,20,50)$ and common covariance $\sigma^{2\prime}=(25,16,400)$. The resultant distribution has been coded with the ordinal values $X_1 \in \{0,\ldots,6\}$, $X_2 \in \{0,\ldots,5\}$ and $X_3 \in \{0,\ldots,4\}$. The transformation was achieved by ranking successive intervals of equal length.

state	pattern	n1	n2	p1	p2
1	010	1	0	0.010	0.000
2	014	1	0	0.010	0.000
3	021	1	0	0.010	0.000
.	.	.	.	.	.
.	.	.	.	.	.
76	633	0	1	0.000	0.010

		Π_1	Π_2	total
size	n_i	100	100	200
	π_i	.500	.500	
mean	1	1.960	3.010	2.485
	2	2.190	3.210	2.700
	3	2.120	2.330	2.225
std	1	1.044	1.124	1.203
	2	.982	.808	1.032
	3	1.047	.975	1.015

$$R_1 = \begin{bmatrix} 1.000 & .014 & .067 \\ & 1.000 & .061 \\ & & 1.000 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 1.000 & -.063 & -.027 \\ & 1.000 & .051 \\ & & 1.000 \end{bmatrix}$$

$$R = \begin{bmatrix} 1.000 & .164 & .068 \\ & 1.000 & .098 \\ & & 1.000 \end{bmatrix}$$

loglinear analysis	significant effects
main effects	X_1, X_2
interactions	-

Table 11.2-8: Characteristics of *NORMAL01* data

11.2.7 *NORMAL02* data

This dataset was generated by discretising a tri-variate normal with original means and common covariance identical to those for the *NORMAL01* dataset. The resultant distribution however was coded with the values $X_1 \in \{0,8,16\}$, $X_2 \in \{0,10,20\}$ and $X_3 \in \{0,40,80\}$ designed to reflect the differences in spread of the X variables.

state	pattern	n1	n2	p1	p2
1	000000	1	0	0.025	0.000
2	000040	2	0	0.050	0.000
3	001000	4	1	0.100	0.025
.	.	.	.	.	.
.	.	.	.	.	.
19	162040	1	5	0.025	0.125

		Π_1	Π_2	total
size	n_i	40	40	80
	π_i	.500	.500	
mean	1	7.400	11.000	9.200
	2	9.000	13.750	11.375
	3	28.000	32.000	30.000
std	1	5.839	5.340	5.847
	2	4.961	4.903	5.453
	3	25.939	22.555	24.235

$$R_1 = \begin{pmatrix} 1.000 & .052 & .070 \\ & 1.000 & -.124 \\ & & 1.000 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1.000 & .014 & -.009 \\ & 1.000 & -.073 \\ & & 1.000 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.000 & .151 & .070 \\ & 1.000 & -.030 \\ & & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	X_2
interactions	-

Table 11.2-9: Characteristics of *NORMAL02* data

11.2.8 *NORMAL03* data

This dataset was generated by discretising a bi-variate normal with original means $\mu_1'=(50,20)$, $\mu_2'=(30,40)$ and common covariance $\sigma^{2\prime}=(100,25)$. The resultant distribution has been coded with the values $X_1 \in \{0,5,10,\ldots,75\}$ and $X_2 \in \{6,9,12,\ldots,42\}$ again designed to reflect differences in spread of the X variables.

state	pattern	n1	n2	p1	p2
1	0024	0	1	0.000	0.001
2	0036	0	3	0.000	0.003
3	0524	0	2	0.000	0.002
.	.	.	.	.	.
.	.	.	.	.	.
137	7518	1	0	0.001	0.000

		Π_1	Π_2	total
size	n_i	1000	1000	2000
	π_i	.500	.500	
mean	1	45.260	27.875	36.568
	2	13.029	29.652	21.341
std	1	9.337	9.365	12.767
	2	4.301	4.260	9.350

$$R_1 = \begin{pmatrix} 1.000 & -.030 \\ & 1.000 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1.000 & .002 \\ & 1.000 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.000 & -.451 \\ & 1.000 \end{pmatrix}$$

loglinear analysis	significant effects
main effects	-
interactions	-

Table 11.2-10: Characteristics of *NORMAL03* data

11.2.9 *POISSON* data

Variable X_1 coded ordinally from *0* through to *11* and X_2 coded ordinally from *0* through to *10*.

state	pattern	n1	n2	p1	p2
1	*0001*	*4*	*0*	*0.004*	*0.000*
2	*0002*	*17*	*0*	*0.017*	*0.000*
3	*0003*	*16*	*0*	*0.016*	*0.000*
.	.	.	.	.	.
.	.	.	.	.	.
108	*1105*	*1*	*0*	*0.001*	*0.000*

		Π_1	Π_2	total
size	n_i	*1000*	*1000*	*2000*
	π_i	*.500*	*.500*	
mean	*1*	*3.020*	*6.981*	*5.001*
	2	*3.042*	*6.958*	*5.000*
std	*1*	*1.737*	*1.733*	*2.633*
	2	*1.778*	*1.778*	*2.645*

$$R_1 = \begin{bmatrix} 1.000 & -.013 \\ & 1.000 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 1.000 & -.013 \\ & 1.000 \end{bmatrix}$$

$$R = \begin{bmatrix} 1.000 & .409 \\ & 1.000 \end{bmatrix}$$

loglinear analysis	significant effects
main effects	-
interactions	-

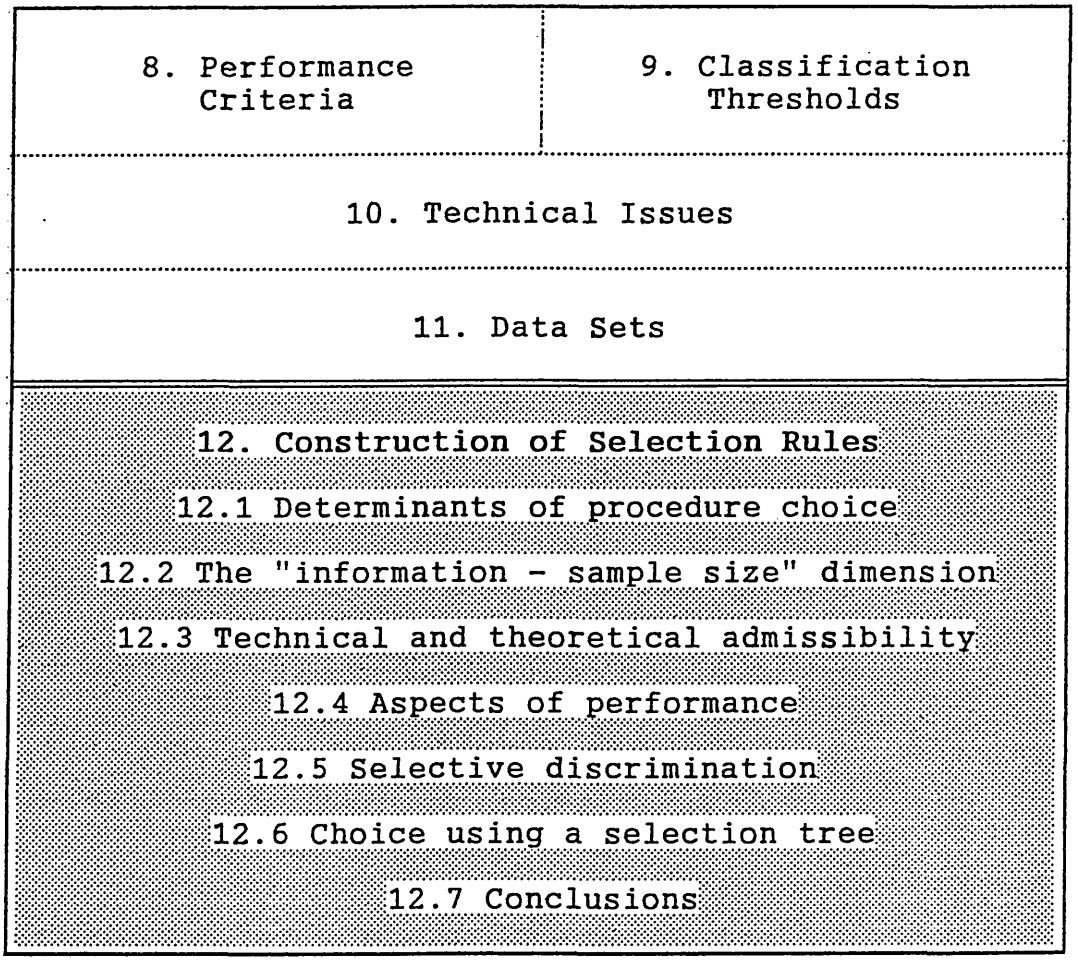
Table 11.2-11: Characteristics of *POISSON* data

The real and artificial datasets illustrated in sections 11.1 and 11.2 exhibit a wide range of characteristics of predictors (dichotomous, nominal, ordinal), dependent population membership variable (*2* populations, *3* and more populations, ordered populations), interactive effects in the predictor variables (none, some, all significant as in the case of the *MA4353* series), prior structure (equal priors, unequal, populations with very small priors), number of discrete states (few, many) and sample size (small, medium, large). Together with the selection rule to be developed in chapter 12 and in conjunction with the analyses in chapters 13 to 15 this material provides a variety of typical examples that researchers may consult when seeking a solution to a particular problem. It must, however, be emphasised at this point that the discriminant ultimately chosen for a new discriminant problem will largely depend on the individual situation. The chances of hitting on a data set with identical characteristics to any of the ones in this chapter are of course remote. Instead the given examples are intended to illustrate the selection process and thus to suggest solutions to a particular problem. For this purpose the range of characteristics of discrete datasets is considered to be adequately representative.

I: INTRODUCTION

II: REVIEW

III: METHOD



IV: RESULTS

V: DISCUSSION

In section 12.1 the *5* key factors that determine choice of procedure are identified: data, demands, constraints, model and skill or experience. These are discussed in detail and the section ends with a brief summary of the major considerations. Of the *5* factors *2* stand out: sample size and prior knowledge about the underlying distributions. These *2* factors lend themselves more readily to formalisation. They are discussed separately in section 12.2. Frequently the technical requirements will allow execution of a discriminant procedure although there may be little theoretical justification for doing so. The near to ubiquitous use of the linear discriminant (chapter 4) for many non-normal situations is a case in point. In section 12.3 the implications of *theoretical admissibility* and *technical admissibility* are contrasted. A catalogue of procedures is developed in terms of either form of admissibility. In section 12.4 the different performance criteria are summarised and placed in the context of the iterative process of procedure selection. Section 12.5 addresses the important question of when it may be advisable *not to* perform a discriminant analysis. Instead a compromise of *selective discrimination* is proposed. This is illustrated by means of a real data example. In section 12.6 the *selection tree* for choice of optimal discriminant procedures is developed. The suggested selection tree is contrasted with other *formal*, *classical* and *economical* approaches to procedure selection. The final tree also includes subtrees for the choice of density estimation and crossvalidation techniques.

12.1 Determinants of procedure choice

The fact that while on one hand there is evidently an abundance of procedures for discriminant analysis (see chapter 4) but on the other hand little comprehensive guidance about how to select any one in a particular

situation (see chapter 3) begs the question as to what has so far guided choice. What are the key considerations that lead eventually to selection of a procedure believed optimal in some sense ?

Five major factors of particular relevance to discrete data situations may be identified in figure 12.1-1.

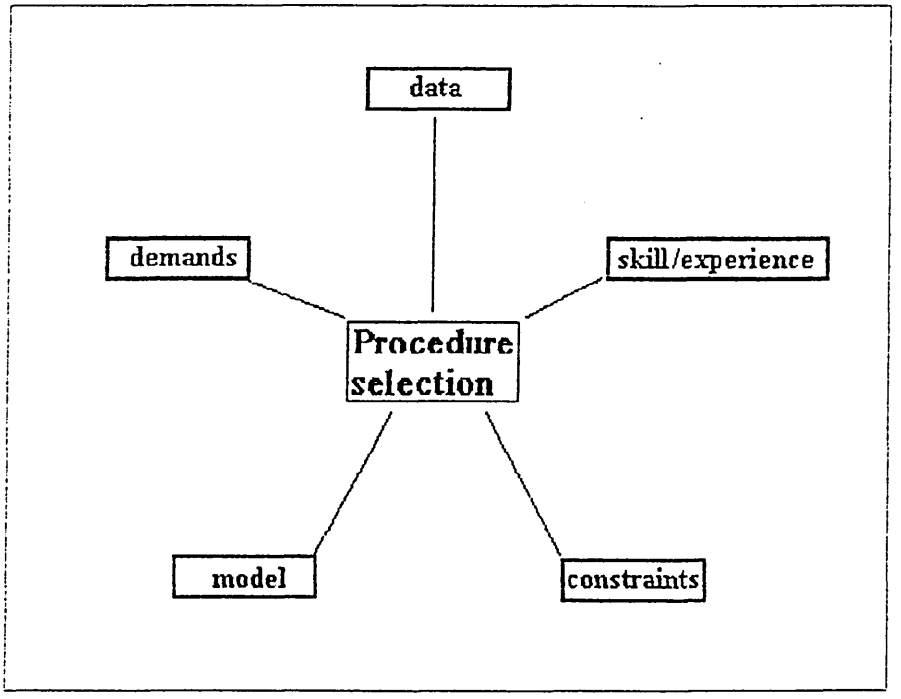


Figure 12.1-1: Determinants of procedure choice

Figure 12.1-1 shows five factors influencing choice of discriminant procedure: (1) amount of data available, i.e., sample size and number of groups, (2) demands placed on the discriminant, i.e., precision, reliability and misallocation costs, (3) general constraints such as *CPU* time, ease of application and availability of procedure, (4) model information, i.e. prior information about the statistical distribution of the data and (5) individual skill and experience of the user in applying the procedure, e.g. setting of initial parameter values and ability of comprehension. These key determinants of procedure choice are discussed in the following.

The sample size available is critical as generally the requirements for crossvalidation vary inversely with the sample size. Larger samples will allow differentiation between more complex models as these typically have more parameters which need estimation. Also stability of estimates is dependent on sample size. On the other hand the gains from a given sample size will be balanced by the number of populations among which discrimination is to be carried out.

There is no hard and fast rule as to how large a sample needs to be in order to support adequate estimation of model *M*, say, nor is there a simple rule of thumb for saying how large a sample must be such that a given discriminant procedure *P* may be conducted.

Next the nature of variables themselves is important. Some discriminant procedures are specifically designed for multivariate binary data such as the parametric Bahadur expansions (see chapter 4 section 1), others for discrete data with small cell frequencies such as the non-parametric procedure based on Hills' concept of optimal interpopulation distance measures (see chapter 6 section 1). Specifically number of populations, type of variable (dichotomous, nominal, ordinal) and the number of categories per variable will be important indicators for the suitability of a given discriminant.

Other aspects of the data factor are missing observations and ease with which further observations may be obtained. A full treatment of missing data estimation has been excluded from the present thesis for reasons given in chapter 3. For completeness, however, it is mentioned at this point. The relevance of missing data on procedure selection is seen immediately from the fact that some newer procedures explicitly allow for missing data while others generally leave handling of missing data to the experience or skill

of the user. The recursive partitioning procedures (chapter 4) for instance enable the user to specify a missing observation for discrete data by simply treating it as if it were another value of the variable.

It is perhaps an inherent desire for perfection that will let any dataset ultimately appear insufficient, no matter how large. Thus a researcher or routine user of discriminant analysis will generally want larger samples. Some data may be cheap to obtain whilst other data may have involved much preparation such as the information gathered in the process of developing a new treatment for a particular disease. In such a situation it may well be reasonable to opt for a limited choice of procedures. This has the benefit of rapid and less costly results compared with obtaining possibly more reliable discriminants at higher costs and perhaps even at the price of not being of benefit any more to the patients because of delayed results.

12.1.2 Demands

The purpose of constructing a discriminant varies with the demands placed on it. One may require stable estimates of the misallocation error or alternatively minimum errors. Further one may wish costs of misallocation minimised. The resultant discriminant rule may also be required to be interpretable in the sense that classification trees allow for instance.

Choice of procedure will depend on the purpose for which it is to be used. Commonly (i) analysis of underlying structure among the variables and (ii) prediction for the purpose of allocation of future observations are distinguished. The former (i) concerns identification of the vital variables assumed to account mainly for differences between populations and is largely synonymous with selection of variables. Structural analysis also

concerns interrelations among the predictor variables. When (ii) prediction, on the other hand, is the dominant aim, a demand for minimum misallocation errors, minimum variance of estimated misallocation errors or a maximum of interpopulation distance measures suitably defined will have priority. A procedure optimally satisfying purpose (i) need not necessarily be optimally suited to satisfy purpose (ii).

12.1.3 Constraints

Sometimes there will be insufficient *CPU* time available or a lack of experience or skill in applying the procedures. Pragmatic considerations may require a quick yet easy to use and cheap procedure at the cost of perhaps poorer performance. Some procedures, particularly the iterative ones, such as recursive partitioning and neural network based discriminant analysis, but also, to a lesser extent, kernel and nearest neighbour density estimation based procedures, can be extremely computer intensive. They require not only considerable *CPU* time but may also be demanding on the much slower machine dependent input-output processes. In addition when the focus is on crossvalidation for the sake of reliability in performance criteria, the computing costs rise for all procedures. This is especially noticeable when bootstrap methods (see chapter 7) are used[40].

Another aspect of constraints limiting procedure choice is the ease with which a given procedure may be implemented. Classical procedures such as the linear discriminant or the logistic are commonly integrated into statistical packages

---

[40] Some procedures such as kernel density estimation based discriminant analysis require certain parameters to be already derived at the estimation stage using bootstrap methods. In this case bootstrap induced computer time is further increased.

that frequently enable direct generation of results, including a variety of density estimation and crossvalidation options.

These built-in procedures have the advantage that, as they are part of the whole statistical package, the problem of data maintenance, i.e. adding, deleting and editing observations poses little difficulty. Packages use standard dataset formats and frequently offer a variety of transformations to and from other packages. They may also allow fixed and free formatted input of raw data. Regarding running the procedures, modern packages again prove superior because of their multiple window front ends. The user can edit procedure specifications in one window, monitor the progress in a log window, and - often simultaneously - inspect the results in a third one. Less common procedures such as the recursive partitioning ones like *FACT* or *CART* have not yet been routinely integrated into standard packages.

Other procedures such as Dillon and Goldstein's distributional distance (1978) or even the simple centroid procedure have been described in the literature making the algorithm clear. However, the implementation, as in this case, is frequently left to the user because these procedures are not integrated into standard statistical software packages. This poses two problems: either the procedure is compiled using a high level language such as the *PL/1* clone in *SAS*, which is a simple task but yields slow execution, or a lower level language such as *FORTRAN* is used thus giving much faster execution speed[41].

As stand-alone programs they all have their individual file handling, data management and program execution peculiarities. In addition not all programs run under popular user front ends such as *Microsoft Windows*. *PACT*

---

[41] This option was used to achieve faster execution for the extensive comparative analyses.

(Shih, 1994), the sequel to *FACT*, even has its own memory manager, a so called *DOS Extender*. The above clearly shows the considerable range of user-friendliness and is an indication of the relevance of the constraints factor for choice of procedure.

12.1.4 Model

Prior information about the statistical distribution underlying the observed data is vital. When such information is available choice of procedure can be considerably enhanced. Information about the data consists of such aspects as information about the prior probabilities of population membership and - in the parametric situation - information about the distribution of the independent variables and the actual model information. In the case of mixture sampling generally the priors, if unknown, may be estimated from the sample. In the case of separate sampling, independent information is required. Information about the statistical distributions concerns class of variables (discrete, ordinal or continuous), type of distribution (normal, Poisson, negative binomial, etc.) and, if known, information on any relevant distributional parameters. If the type of distribution is unknown a nonparametric procedure may be indicated.

12.1.5 Skill and experience

Last but not least important is the individual user's experience with a given procedure. Judicious selection of appropriate starting values for some of the iterative procedures such as artificial neural nets (see chapter 4) is needed. Skill and experience in application is also needed for detecting important features in the data, particularly interactions, and modelling them appropriately. The user's skill will depend on experience

in the application of discriminant procedures and on professional background; i.e. on whether he/she is a frequent or an occasional user. In this context it is perhaps useful to point out one of the difficulties in conducting so called *meta-analyses* of published research reports. Assume that several reports containing results from discriminant analysis are to be summarised. A spontaneous response might be to average reported misallocation errors. A possible refinement might be to weight individual reported errors by sample size as in meta-analyses. The problem of prediction of premature deliveries is a good example of this situation. The review of studies on prediction of premature deliveries conducted by Shino and Klebanoff (1993) reports positive predictive values ranging from *10%* to *30%*. Such a wide range may be partly due to differences in skill and experience between individual researchers and not entirely due to real differences between samples.

12.1.6 Summary

Figure 12.1-2 summarises the above points. For each major factor influencing choice related aspects are shown. Their order of appearance indicates relative importance. Of course the boundaries between the above five key determinants of procedure selection (data, demands, constraints, model, skill/experience) are not quite as sharp as may appear from figures 12.1-1 and 12.1-2. Consider a user who wishes to regularly extract only a few standard performance statistics. If presented with good information about the data such that he/she can be always sure that the observed feature data $\mathbf{X}_i$ is distributed according to some known distribution $F_i(\mathbf{X})$ with parameter vector θ_i and prior probabilities π_i then such a user will be relieved of additional model selection tasks. Conversely, a skilled and experienced statistician may well be able to extract missing distributional information from

a given sample and thus skill and experience make up for lack of model information.

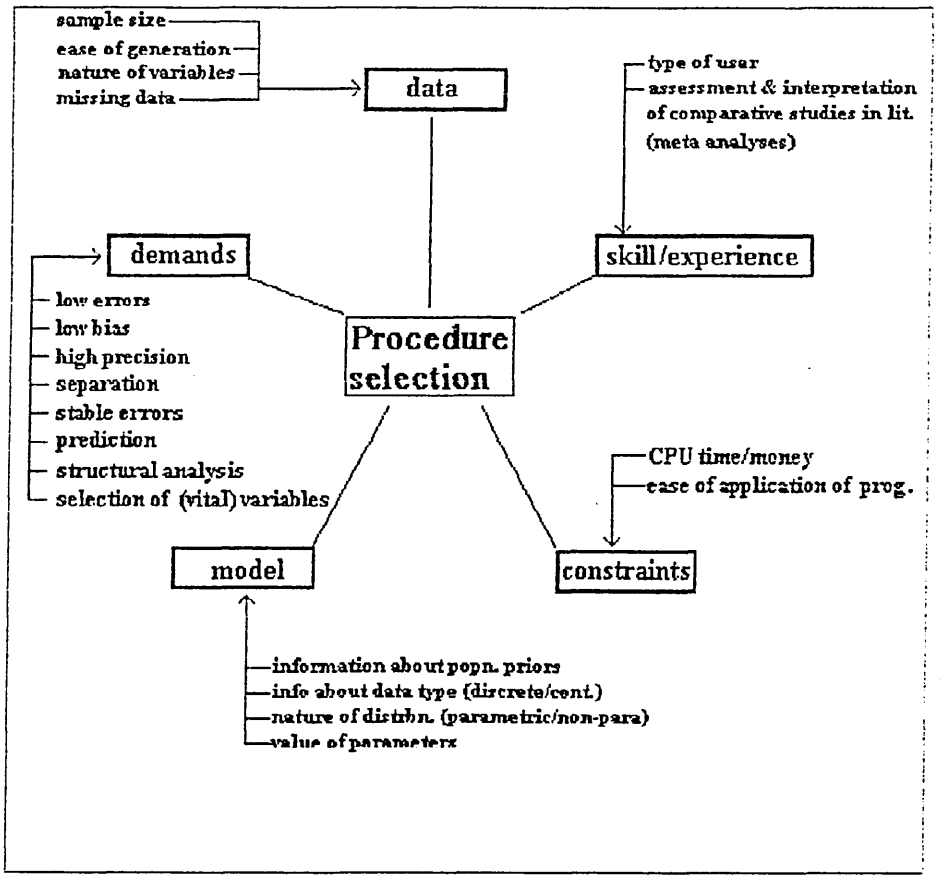


Figure 12.1-2: Detailed selection factors

12.2 The "information - sample size" dimension

To a certain extent some of the factors mentioned above need to be balanced such as demands and constraints. Others may synergise such as skill/experience, data and model. To demonstrate the effects on scope of procedure choice consider first a not uncommon supposition of unconstrained resources, a high level of skill, experience and high demands. Next assume degree of prior model information and sample size to vary independently. This scenario is depicted in figure 12.2-1 which shows scope of procedure choice given prior information. The range of procedures one may wish to select from varies with sample size and prior knowledge about the distribution of the data. When neither

data nor any information about its distribution are available no procedure can be selected. When on the other hand infinite data and maximum information is available the optimal procedure may be selected with certainty. Between these two (unrealistic) extremes a rich variety of selection options exists. Generally when sample size is large yet data information is scant one would opt for more robust procedures. If however, sample size is small yet considerable information about the underlying distribution is available more specific procedures may be chosen. In the ideal situation - remembering that skill, constraints and demand are held constant - sufficient data and comprehensive prior information would be available. Choice of procedure should be an easy task.

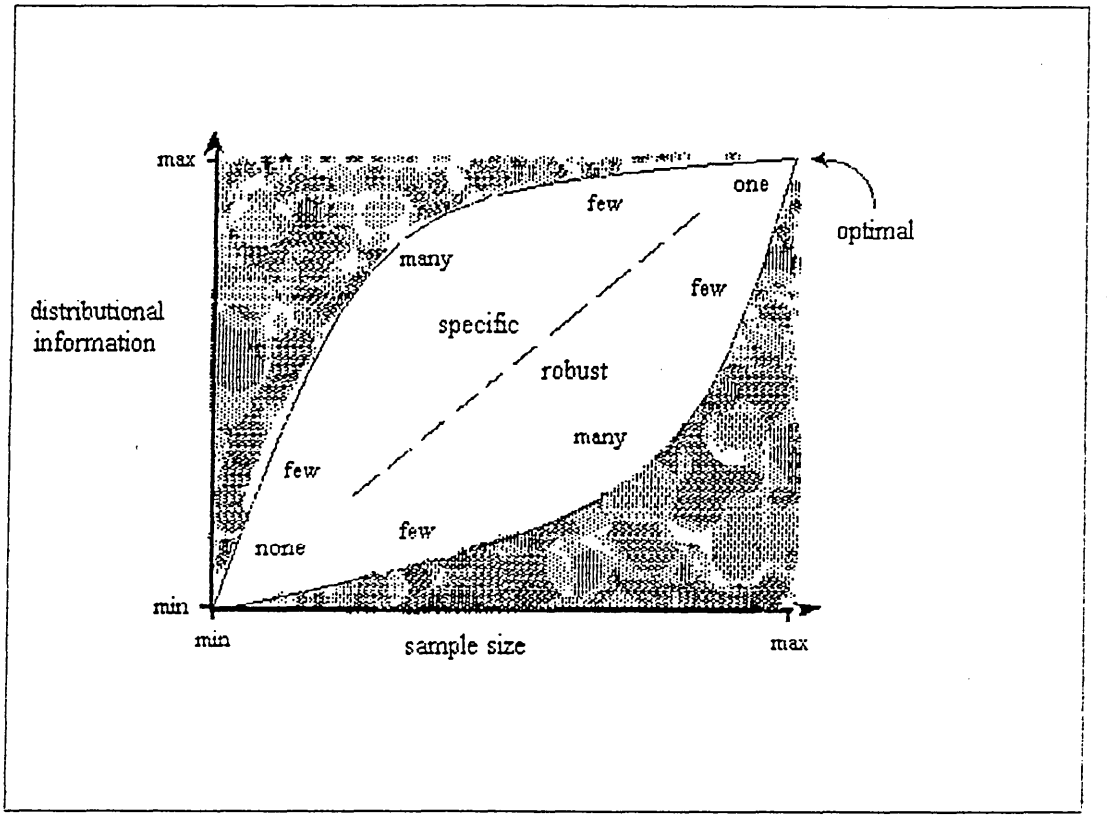


Figure 12.2-1: Scope of procedure choice

In practical situations these conditions are rarely satisfied. Perfect information would allow immediate construction of the discriminant rule and infinite sample size would give maximum precision ($Var\ (\varepsilon) = 0$). Commonly

some prior model information is available as well as data of size $n = \sum_i n_i$. In discrete discriminant analysis frequently the feature variables X_{ijk}, $(i=1,\ldots,g;\ j=1,\ldots,n_i;\ k=1,\ldots,q)$, are bounded i.e., $X_{ijk} \in \{1,2,\ldots,l_{ik}\}$ where l_{ik} is the number of discrete levels of X_{ik}.

Thus often a multinomial data model may be assumed, yet information is incomplete as the parameters of its marginal distributions still require estimation. Further there will rarely be any useful prior information on the interactive structure of the data. If only few data points are available for the estimation process this will imply merely basic models while large sample sizes allow estimation of more complex models. The factors *model* and *sample size* are thus seen to interact: given sufficient sample size additional model information may be gleaned by inspecting the data in support of various models. Obviously, the less data there is - in absence of complete model information - the fewer models there will be to choose from which leads ultimately to the degenerate case of *"no data ⇒ no model ⇒ no discriminant"* (figure 12.2-1).

It is important to realise that with few data samples and little information about data no useful discriminants may be reliably selected. An ideal situation would exist if perfect distributional information were available. In this case we have the exact population specific distribution functions, the exact values of their respective parameters and exact values for the prior probabilities of population membership. From these quantities the posteriors may be derived directly. Thus an optimal parametric discriminant rule may be found even in the absence of any training data. In practical situations however total information is not available. On the other hand assuming large sample sizes and no distributional information we have the classical scenario for nonparametric discriminant analysis. An optimal procedure must be searched for. This requires estimation of the density from the sample data by

nonparametric methods and, because of large sample sizes these estimates will tend to be good.

Both these extreme situations are in practice never attained. Generally some data are at hand while information about underlying distributions may not always be available. Figure 12.2-1 shows examples of a few typical situations in an information by sample size plane. The upper left and lower right corners of the plane are never reached. Instead, depending on whether more information is available or whether more data samples are given the optimal procedure can be approached either by going from few to more predominantly specialised discriminant procedures (upper curve) or by going from few to more predominantly robust discriminant procedures (lower curve).

12.3 Technical and theoretical admissibility

In figure 12.2-1 the joint effect of two of the five determinants of choice, theoretical and empirical information, are illustrated. Theoretical information pertaining to the distribution of the data comprises

(1) a probability model for the distribution of $\mathbf{X}_i$, i.e. $Pr\{\mathbf{X}=\mathbf{x}|\Pi_i\}$,

(2) (estimates of) prior probabilities $\pi_i=Pr\{\mathbf{X}\varepsilon\Pi_i\}$,

(3) specification of the cumulative distribution function

$$F(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(t)dt \text{ and}$$

(4) (estimates of) relevant parameters for $F(\mathbf{x})$.

Empirical information pertaining to the quality of the data comprises

(1) sample size in terms of number of available observations,

(2) presence of missing observations,

(3) availability of correctly labelled objects in the sense of *supervised learning* and

(4) type of sampling frame, i.e. separate versus mixture sampling.

The theoretical and empirical information listed above also applies to continuous discriminant analysis. For completeness the empirical information also includes missing data and labelling of objects. As pointed out in chapter 3 these aspects were deliberately excluded such that complete and correctly labelled data are assumed. Similarly separate sampling frames are assumed for reasons given in chapter 10. Theoretical and empirical considerations alone will generally point to a likely initial candidate for a suitable procedure. Crucial for the selection however is its performance. The relationship between these three determinants of choice is discussed further in section 12.6. Theoretical and empirical information are seen as initial inputs to the selection process while performance is secondary as it follows execution of a given procedure. Performance evaluation is vital as it modifies future selections and thus is seen as a central feature in the iterative process of procedure selection (see section 12.6).

The relevance of theoretical and technical admissibility for the selection process may be demonstrated by compiling a catalogue of procedures and mapping admissibility regions onto a plane defined by type of predictor and type of response. This plane is given in terms of the dimensions of order of independent variable (i.e. dichotomous through to continuous) and order of response (i.e. dichotomous through

to ordinal). Essentially the same ordering is adopted as for tabulating the datasets in chapter 11. Figures 12.3-1 and 12.3-2 show respective regions of admissibility for eight classes of discriminant procedure (linear/quadratic discriminant function, logistic, Bahadur, Lancaster, Hills distance, distributional distance, centroid, and *CART/FACT*). In all diagrams the response is plotted vertically and the predictor is plotted horizontally.

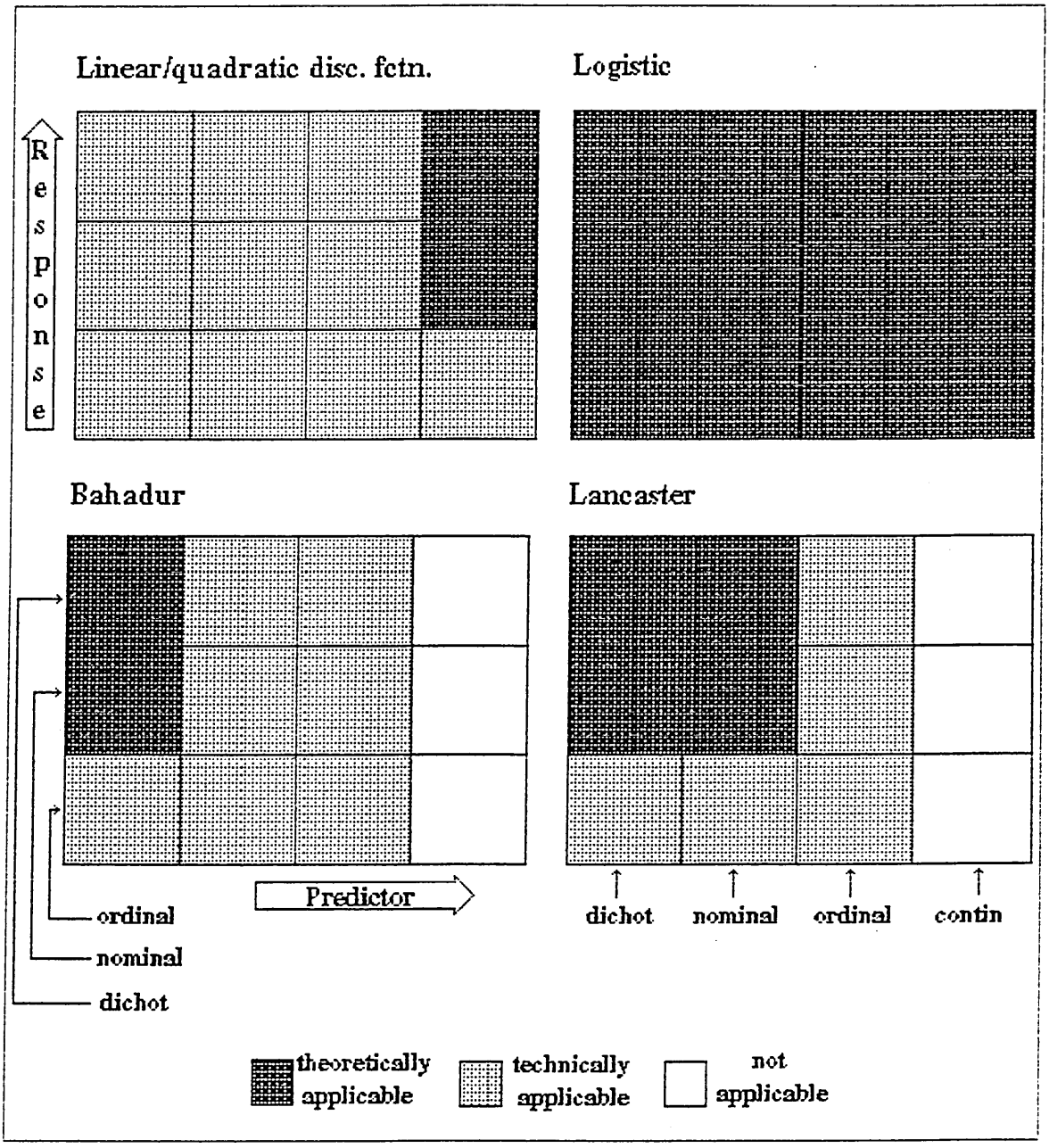


Figure 12.3-1: Catalogue of admissibility regions

The linear and quadratic discriminants require multivariate normal distributions yet they have proved to be widely applicable due to their robustness to departures from the distributional assumptions. In standard applications ordinal responses are treated as nominal (Anderson, 1984). The logistic model - probably the most widely used discriminant procedure for discrete data - stands out because its theoretical admissibility region is coincident with its technical admissibility region. The ordinal response is modelled by using *cumulative logits* while the additional information contained in ordinal predictor variables is utilised by directly entering the variables into the design matrix. The Bahadur allows the joint representation of dichotomous data directly in terms of means of and correlations between independent variables. Thus it is ideally suited for situations in which the data exhibits interactive structure. There is no special modelling of ordinal responses. Lancaster models are equivalent to Bahadur expansions except that they are not restricted to dichotomous predictor variables but can also be applied to nominal data. As with the Bahadur models however there is no special modelling of ordinal responses.

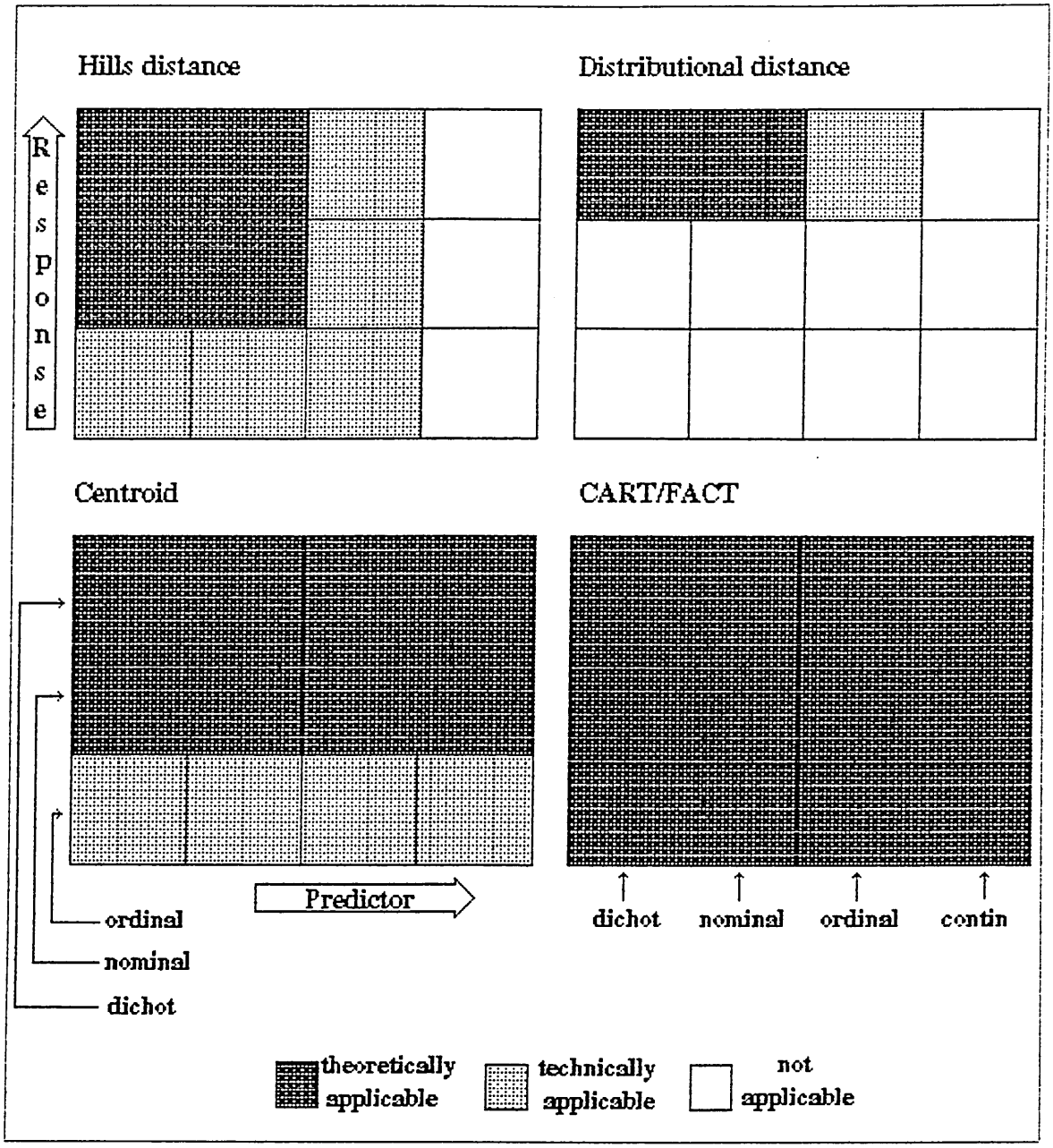


Figure 12.3-2: Catalogue of admissibility regions

The interpopulation distance measure suggested by Hills (1967) can be extended to more than *3* populations (see chapter 10) and is therefore applicable to nominal responses in general. The distance measure is specifically designed to be sensitive to small changes in individual state probabilities. There is no special modelling of ordinal responses. Goldstein and Dillon (1978) developed the generalised distance rule only for a dichotomous response. It is specifically designed for discrete data represented in state matrix notation. The centroid

procedure is easy to apply to any type of predictor variable. Again there is no special modelling of ordinal responses. Recursive partitioning procedures are nonparametric and make no assumptions about the data distribution. It suffices to simply specify the class of variable. Therefore theoretical and technical admissibility regions coincide.

Because theoretical admissibility necessarily requires technical admissibility the region where a procedure "ought to be applied" is always a genuine subset of the region where it "could be applied". Looking through the figures 12.3-1 and 12.3-2 one may read table 12.3-1 directly off the graphs. Another way of jointly presenting the information in a single *3*-dimensional diagram is shown in figure 12.3-3 for technical admissibility and figure 12.3-2 for theoretical admissibility. Comparison between those figures shows immediately the much larger domain for technical admissibility.

		max admissibility region	
procedure	variable	technical	theoretical
distributional distance	response predictor	dichotomous ordinal	dichotomous dichotomous
Bahadur	response predictor	ordinal dichotomous	nominal dichotomous
Lancaster	response predictor	ordinal ordinal	nominal nominal
Hills distance	response predictor	ordinal ordinal	nominal nominal
Centroid	response predictor	ordinal continuous	nominal continuous
recursive partitioning CART/FACT/CHAID	response predictor	ordinal continuous	ordinal continuous
logistic	response predictor	ordinal ordinal	ordinal ordinal
linear/quadratic discriminant function	response predictor	ordinal continuous	nominal only continuous

Table 12.3-1. Summary of admissibility regions

Table 12.3-1 shows maximum technical and theoretical admissibility regions of discriminant procedures for response and predictor variable. The theoretical admissibility region necessarily is always a genuine subset of the technical admissibility region.

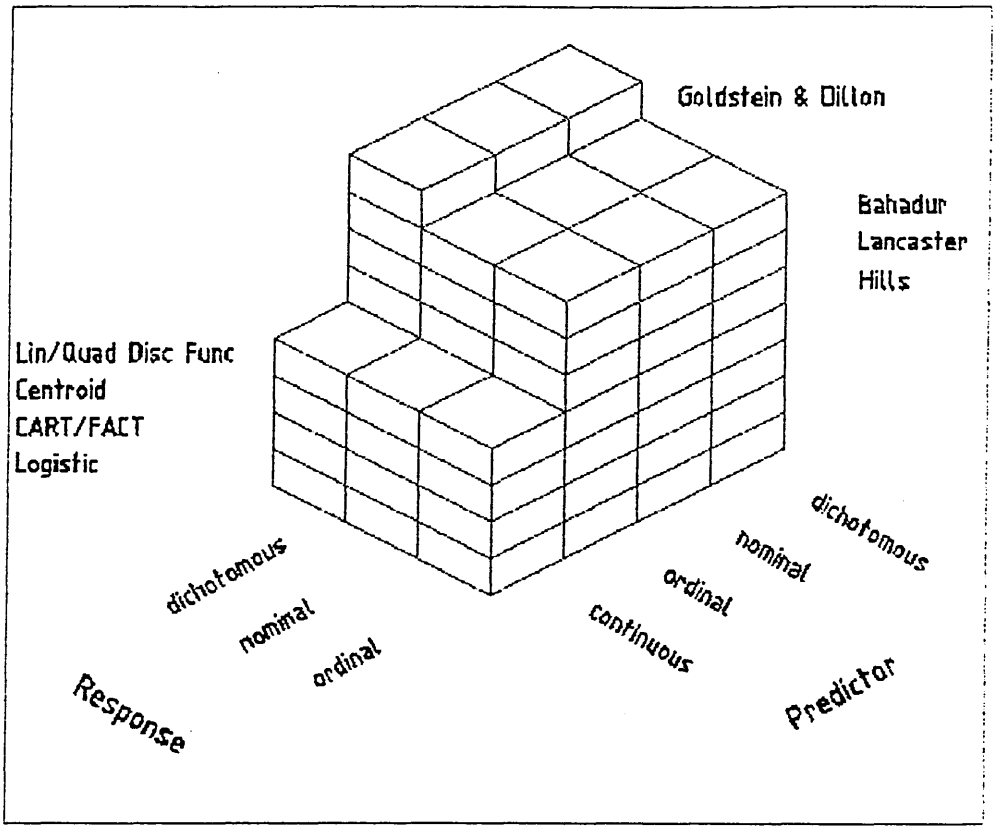


Figure 12.3-3: Technical admissibility

Figure 12.3-3 shows regions of technical admissibility presented jointly for *8* discriminant procedures. All except the Bahadur procedure may be used in situations where the response variable *Y* is ordinal. The Bahadur is also very stringent as it is further restricted to dichotomous independent variables.

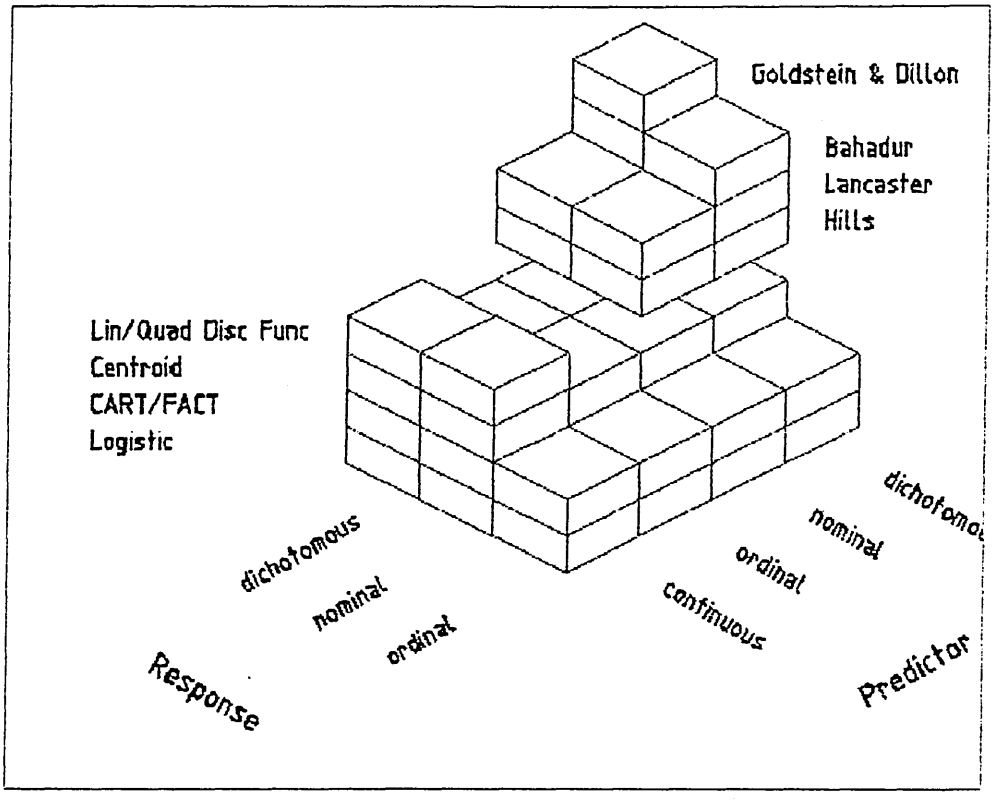


Figure 12.3-4: Theoretical admissibility

Figure 12.3-4 shows regions of theoretical admissibility presented jointly for *8* discriminant procedures. This is clearly a subset of the admissibility regions shown in figure 12.3-3. Only *2* procedures allow specific modelling of ordinal responses. The linear discriminant although widely applied is theoretically only admissible for continuous data.

Figures 12.3-3 and 12.3-4 show regions of technical and theoretical admissibility for all eight procedure classes jointly in one diagram. The comparison immediately reveals the drastic reduction in range when only theoretical admissibility is considered. The limited range of theoretical admissibility for the classical linear/quadratic procedure is due to the stipulation of normality and its implicit continuous data structure. This example shows how relative the demands for a purist approach to the discriminant problem can be. Therefore when dealing with selection of a suitable procedure the option

of choosing technically applicable yet theoretically inappropriate procedures must be clear from the start.

12.4 Aspects of performance

As indicated in 12.3 performance follows initial selection on the basis of model information and information on the data. From chapter 7 on performance evaluation it was seen that four aspects may be distinguished: (1) error rates or equivalently hit rates, (2) separation between the populations, (3) bias reduction and (4) reliability. These demands outlined in section 12.1.2 break down further as follows:

(1) misallocation errors or hit rates

- the "classical" counting based estimates such as $\varepsilon_{counting}$
- posterior probability based estimates such as $\varepsilon_{posterior}$ (Hora and Wilcox, 1982) and *ALS* or *AQS* (Titterington et al, 1981)
- smoothed error rate estimates such as Glick (1976)

(2) separation measures

- based on distances such as
  - "classical" Mahabanobis distance D^2
  - standard Euclidean metric
  - interpopulation distance measure of Hills (1967) & modified by Lack in the context of the present research
  - Goldstein and Dillon's (1984) generalised distributional distance based on Matusita's (1955) population distance

- based on functions of posterior heterogeneity such as
  - classical direct η criterion (Lack)
  - indirect η criterion using pseudo posteriors (Lack)

(3) bias reduction or (cross)validation methods

- hold-out, or separate training and test sets
- leave-one-out (Lachenbruch, 1975)
- resubstitution
- hold-v-out, where $1 < v < n$
- conditional bootstrap
- unconditional bootstrap

(4) reliability measures

- receiver operating characteristics (*ROC*) eg Cole et al (1991)
- *DT* score of Titterington (1981)
- constant thresholding of posteriors as implemented for instance in *PROC DISCRIM* of *SAS* version 6.3
- variable thresholding for *3* or more groups (Lack) as formulated in chapter 9

The above list demonstrates the breadth of tools available for performance assessment. Figure 12.4-1 shows the major role played by performance evaluation in the iterative phase of the selection process.

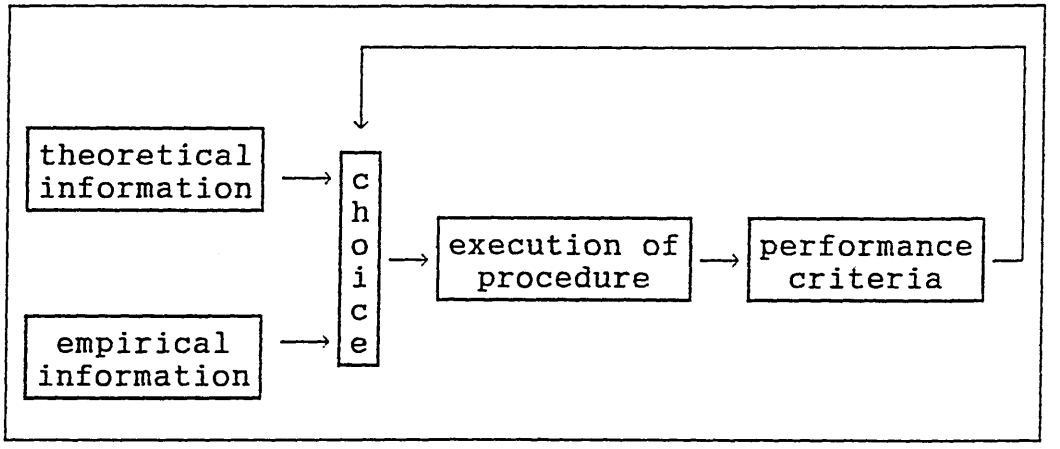


Figure 12.4-1: Iterative selection process

Figure 12.4-1 shows theoretical and empirical information guide choice initially. Subsequently a major role is played by performance evaluation. The feedback on the choice stage is seen as an iterative process.

Guides to selection of optimal rules, therefore, will have to not only prescribe the first phase in figure 12.4-1, i.e. choice on basis of empirical and theoretical information, but also indicate appropriate criteria for performance evaluation in the iterative second phase.

12.5 Selective discrimination

Before moving on to construction of the final tree selection, the issue of *when not to perform discriminant analysis* needs addressing. It is not uncommon that discriminants urgently required for crucial problems show little promise of reasonably low misallocation rates. One of the examples in chapter 2 illustrated the dilemma faced when the distributions of different populations overlap to a considerable extent. A typical example of data with large overlap, and thus poor predictive power occurs in the prediction of preterm delivery. Two classes of influencing factors may be distinguished: medical and psychosocial. The non-medical determinants of early birth onset are not all

clearly defined and measurable. Medical factors believed associated are:

- previous preterm delivery
- previous history of abortions
- previous perinatal death
- past birth complications
- previous caesarean section
- past uterine surgery
- short interpregnancy interval
- bleeding in third trimester
- placental rupture
- placental insufficiency
- premature labour
- anaemia
- hypertension
- hypotension

Psychosocial factors believed associated with preterm delivery are:

- reported difficulty in coping with a job while pregnant
- living without a supporting spouse
- lower socioeconomic status
- consumption of more than *10* cigarettes per day
- stress and tension arising within the family
- stress and tension arising at work

The data presented is extracted from routinely gathered information in the Bavarian perinatal survey (Bayerische Perinatalerhebung, *BPE*)[42]. The items listed below are a

---

[42] Preliminary analyses investigating whether useful predictors can be found for premature deliveries were recently published in the annual report on the Bavarian Perinatal Survey, BPE, by Lack (1994). The data analyzed relates to *112246* singleton Bavarian deliveries during 1993. The main findings indicate comparatively small odds ratios suggesting that prediction of premature deliveries is prone to errors.

subset of the standard data form completed for every birth. The data presented in table 12.5-1 relates to singleton pregnancies with the dependent variable *premature birth* defined as gestational ages less than *37* completed weeks of pregnancy. In the analysis up to *17* different medical factors and *8* different psychosocial factors were considered. To reduce the number of cells to a useful amount two variables were derived: counts of number of psychosocial factors and number of medical factors recorded at birth. These counts were further given ceilings of *2* and *3* respectively. Inspection of the table shows that mature births dominate every one of the *12* possible cells. The proportion of premature births is always lower than that of mature deliveries in all cells. This is largely accounted for by the low prevalence of about *7* percent. This is a realistic example in as much as factors considered relevant reveal little predictive potential even when considered jointly. The predictors are however not completely useless as the last column shows. The probability of premature delivery increases with the number of reported factors.

no of psychosoc factors	no of medical factors	prem births	mature births	joint density	probability of prem birth
0	*0*	*186*	*5142*	*0.530*	*3.5*
0	*1*	*161*	*1183*	*0.134*	*12.0*
0	*2*	*48*	*198*	*0.024*	*19.5*
0	*3*	*22*	*52*	*0.007*	*29.7*
1	*0*	*58*	*1815*	*0.186*	*3.1*
1	*1*	*66*	*449*	*0.051*	*12.8*
1	*2*	*24*	*81*	*0.010*	*22.9*
1	*3*	*6*	*15*	*0.002*	*28.6*
2	*0*	*25*	*388*	*0.041*	*6.1*
2	*1*	*21*	*74*	*0.009*	*22.1*
2	*2*	*2*	*18*	*0.002*	*10.0*
2	*3*	*5*	*8*	*0.001*	*38.5*
		624	*9423*	*1.000*	

Table 12.5-1: Prediction of preterm delivery

Table 12.5-1 shows a contingency table of the number of premature and mature deliveries by number of medical and

psychosocial factors believed causally related to premature delivery. The penultimate column shows joint density estimates for both populations[43]. The final column gives the probability of premature birth computed from the raw cell counts[44]. This increases with number of recorded factors from *3.5%* for no factors present to *38.5%* for the maximum number of factors present.

It would be of considerable value for the maternal and child health services if a reasonable predictor of premature delivery were available. With it health care resources could be more efficiently allocated while expenses required for the care of children with handicaps resulting from premature birth could be lowered. Perinatal traumas and postnatal defects could be reduced. But inspection of the data in table 12.5-1 reveals that for the majority of observations there is little to choose from among all possible discriminants. The probabilities of premature delivery increase to sizes that may be of use in practical settings only towards the tails of the joint distributions for groups Π_1 and Π_2. This relationship is illustrated in figure 12.5-1 where the last column of table 12.5-1 has been plotted against the last column but one.

---

[43] E.g. *(186 + 5142) / (624 + 9423) = 0.530* .
[44] E.g. *186 / (186+5142) * 100 = 3.5* %.

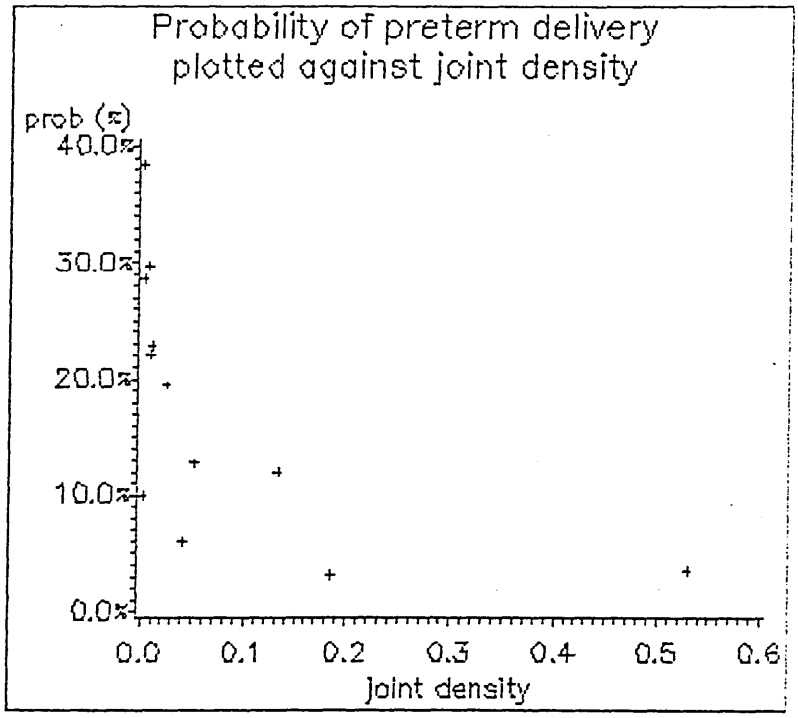


Figure 12.5-1: Probability of premature birth

The example illustrates the problem of deciding when discrimination analysis is sensible and when it should better not be performed. A natural answer would be to refrain from discriminant analysis when the performance is expected to be low. However such a criterion will not suffice for all selections as shown above. In these situations it may prove useful to concentrate on only part of the data.

The idea behind such *selective discrimination* is essentially pragmatic and consists in restricting prediction only to a selected subsample of the population where the feature vectors are extreme enough to support reasonable results. By doing this at least some of the more extreme cases of prematurity may be identified with acceptable specificity. *Selective discrimination* of course raises the question of where to segment off the data one wishes to use and once it has been segmented off again the question will be how to select the optimal procedure. This problem will not be pursued because the segmentation will always have to take into account issues particular to the

problem. For instance in the present example the costs incurred by following up false positives will be of major concern leading to a reduced sample. On the other hand the gain to be expected from intervention in the true positive group will be an argument for increasing the sample size.

12.6 Choice using a selection tree

The requirement stated at the end of section 12.4 that guides to selection should also suggest appropriate performance criteria for evaluation, highlights the dilemma of choice: performance criteria are both crucial determinants of the selection process and also a yardstick for procedure evaluation.

An example may illustrate this further. One may obtain higher hit rates with discriminant procedure *P* but this may be of little use if the demands (see figure 12.1-2) require a greater emphasis on reliability or low variance. To put it pointedly: "all desirable qualities may be correlated to a certain extent but will rarely coincide". Once choice has fallen on a particular performance criterion, φ say, the selection process simplifies (figure 12.6-1).

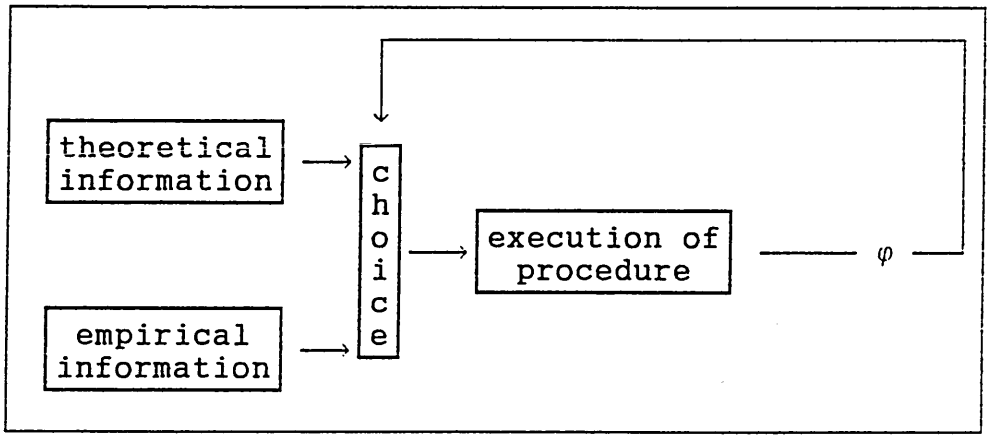


Figure 12.6-1: Simplified selection process

Figure 12.6-1 shows a simplified selection process with a fixed performance criterion φ say. The iterative second stage in the selection process still remains but is substantially simplified and therefore faster.

The difference between figures 12.4-1 and 12.6-1 is that the former gives an overview of all possible determinants of choice while the latter demonstrates a common paradigm used when the question about the type of performance criterion has been settled. This points to a crucial stage in the selection process, namely deciding, preferably at an early stage, on the yardstick by which performance should be assessed - and then sticking to it ideally. Consequently priority should be given to answering the question *"What do I want the discriminant for ? What do I want it to do well ?"* Clearly committing oneself to a simple performance criterion will lead to more narrowly defined situations.

To illustrate the joint effect of theoretical information, information about the data and calculated values of performance criteria consider the following hypothetical situation. Prior information suggests in advance that a given dataset is distributed according to a 3^{rd} order Bahadur expansion, i.e. the sample consists of multivariate binary observations $\mathbf{x}$ whose distribution function $F(\mathbf{x})$ is characterised by interactions between up to three variables at a time. However, if the sample size is small (eg $n = 40$ observations) one will stand little chance of finding a reliable discriminant based on the third order Bahadur model. The sample is too small to provide stable estimates for the means and correlations. It is well known that in such situations the linear discriminant by requiring fewer parameters and by virtue of its general robustness, will provide quite adequate results (McLachlan, 1992). The third order Bahadur model is *theoretically applicable* given the prior distributional information. Considering the small sample size, however, it is advantageous in the example to use the *technically applicable* linear discriminant. The fact that theoretically inappropriate procedures may lead

to more favourable performance than the "proper" procedure has been lucidly shown two decades ago by Victor (1976) (see chapter 7).

Traditionally choice of discriminant procedure assumes distributional information is available and therefore begins with parameter estimation of the conditional distributions for a given model. This approach contrasts with a more formal and general approach where, in the absence of such information, the entire range of available procedures is inspected for technical admissibility. The next stage of the formal approach consists of narrowing the choice down on the basis of further data characteristics. In the last stage selection is made on the basis of performance of the discriminant rule. These different approaches are illustrated in figure 12.6-2.

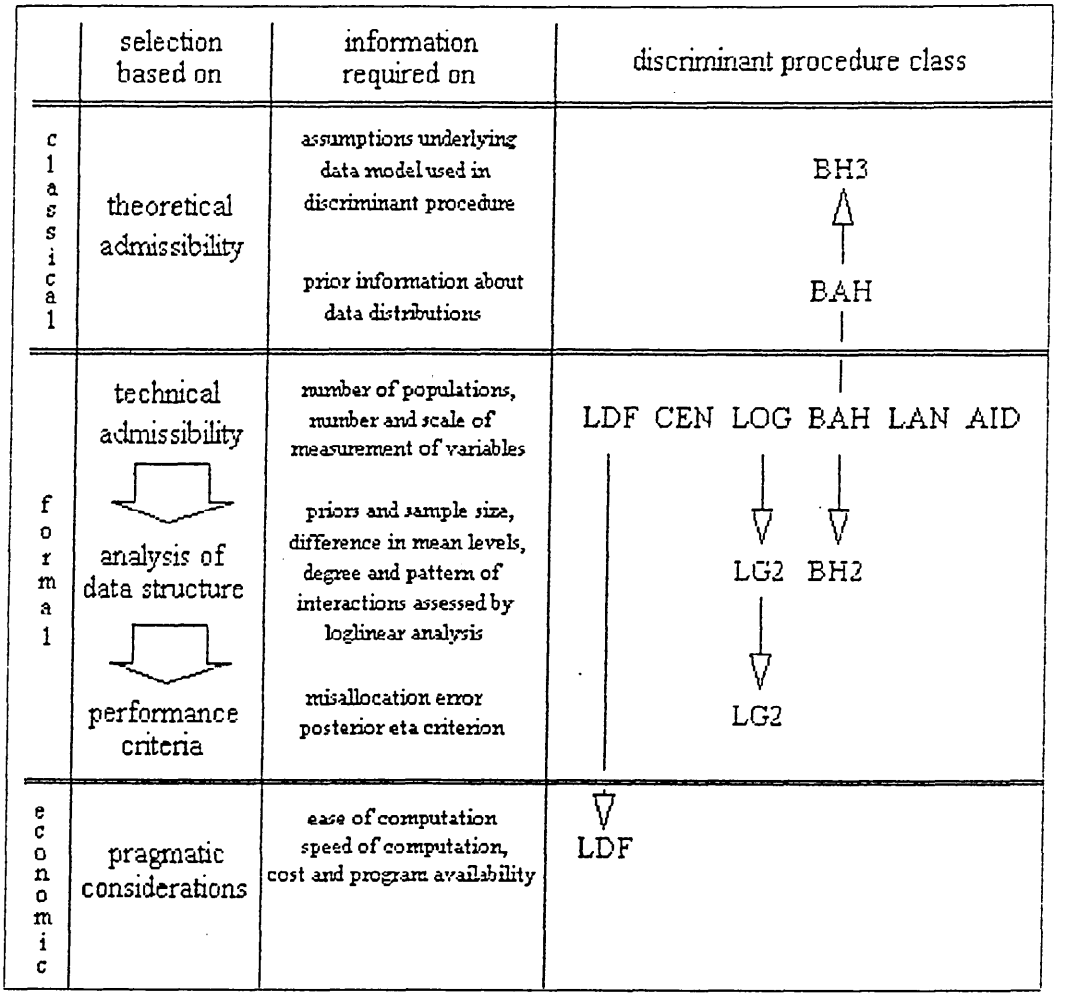


Figure 12.6-2: "Classical" versus formal selection

Figure 12.6-2 contrasts the classical and the formal procedure selection paradigm. The *classical* selection approach *in which the data model is assumed to be known* is based entirely on theoretical considerations. In the above hypothetical example assume that background information about a data sample and its underlying distribution suggest initially a Bahadur model (*BAH* in the diagram). Further information suggests a third order model (*BH3*). The *formal* approach by contrast initially admits all procedures, then narrows choice down to either a logistic or Bahadur model of order *2* and finally homes in on the logistic because of a lower error rate. *Pragmatic* considerations on the other hand may lead to an entirely different model, in this case the linear discriminant function which is robust and readily available in most statistical packages.

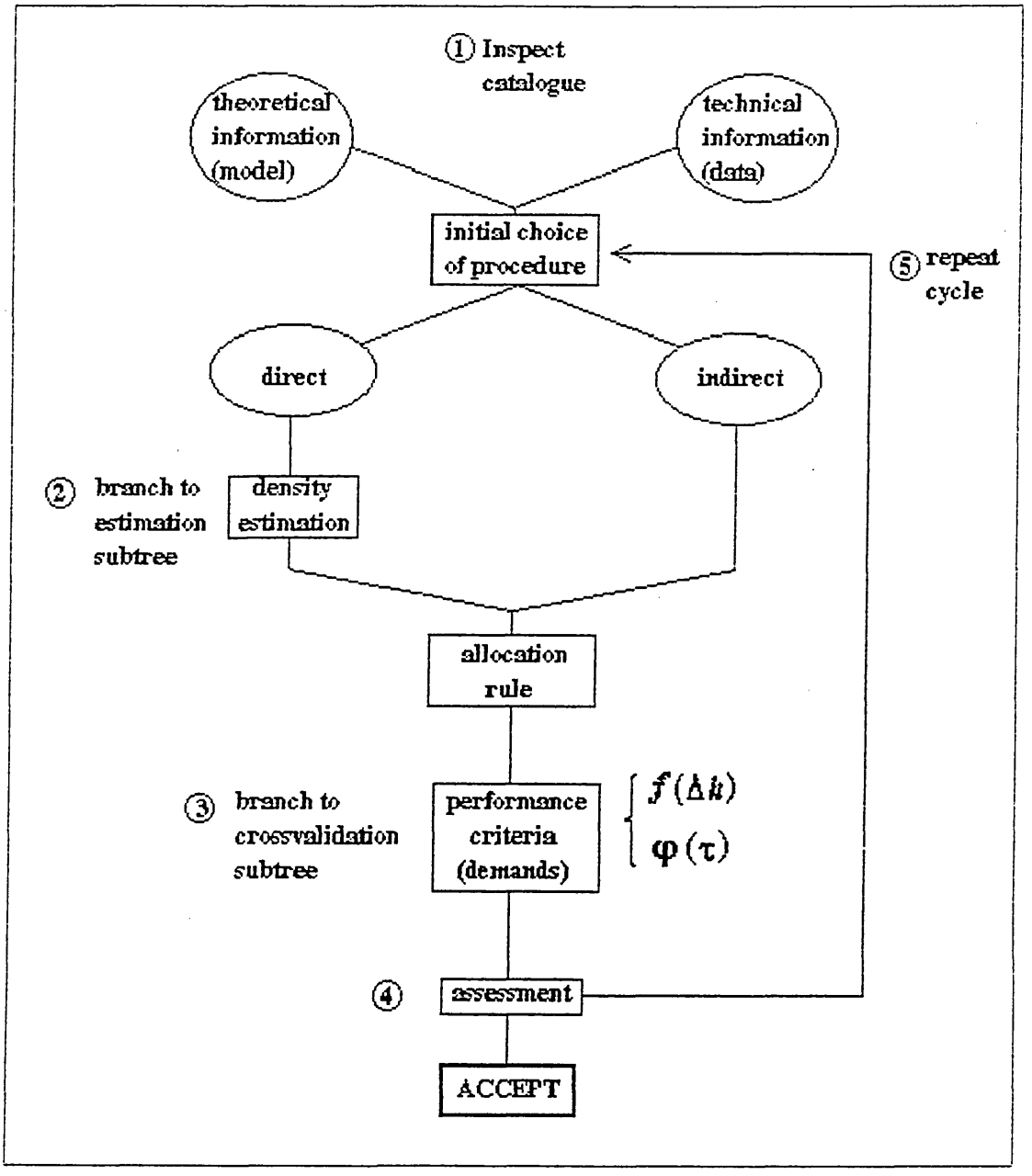


Figure 12.6-3: Procedure selection tree

Figure 12.6-3 summarises the complete selection tree for choice of discriminant procedure for discrete data situations. Numbers in the figure are referred to in the text.

The tree begins at the top where a catalogue of procedures in terms of theoretical and technical admissibility is inspected like a checklist in an almost mechanical sense (1). Initial procedure selection is carried out in two stages. First technical applicability is checked, then

theoretical and empirical considerations narrow choice down further. The latter is done by looking at four aspects in turn: theory and model information, sample size, inspection for possible interactions and metric of the data. In the case of initial choice of a direct procedure a technique for density estimation has to be selected (2). This is represented by the sub-tree shown in figure 12.6-4 illustrating options for parametric and nonparametric density estimation techniques.

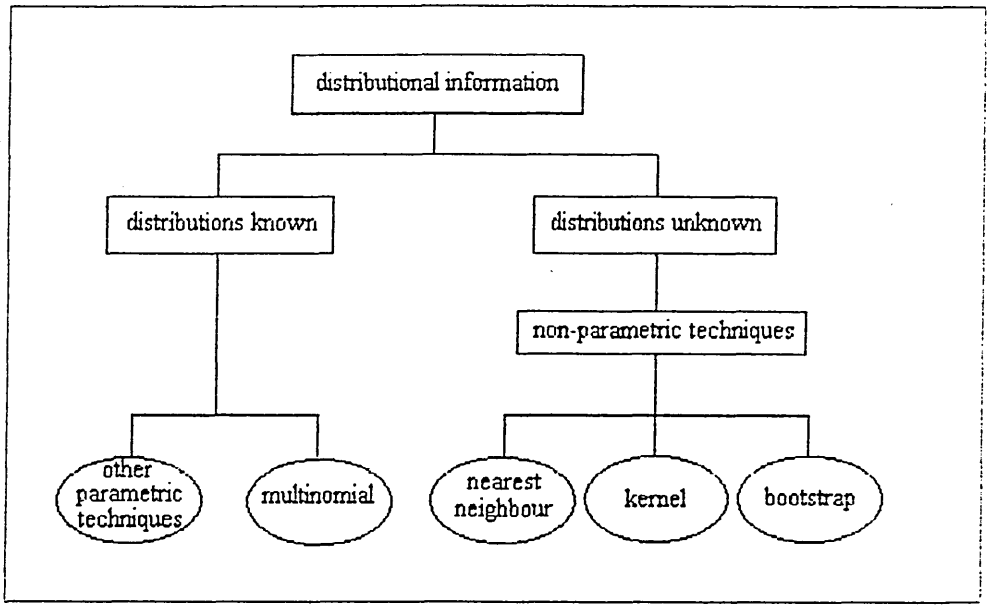


Figure 12.6-4: Density estimation decision tree

The allocation rule is next computed according to the algorithm for the chosen procedure. The demands placed on the discriminant procedure will point to suitable performance criteria. Choice will be among the the error rate estimators $\varepsilon^{counting}$ and $\varepsilon^{posterior}$ as well as the criterion, η. In addition the empirical distribution of relative differences, $f(\tau)$, and the threshold related performance curves, $\varphi(\tau)$, will play a central role here. Estimation of these performance criteria will in turn require a suitable crossvalidation technique (3). Considerations for choice are shown in figure 12.6-5. The technicalities of estimation of performance criteria as

well as the estimation of direct and indirect discriminant procedures are outlined in chapter 10, section 5. If the performance criteria meet the demands a chosen discriminant procedure may be accepted after the assessment stage (4). Otherwise the initial choice of procedure may be modified (5) and the iterative process of procedure selection enters a second cycle. If all chosen discriminant procedures are so deficient that even the minimum acceptable performance is not met the question of *selective discrimination* (section 12.5) might be considered.

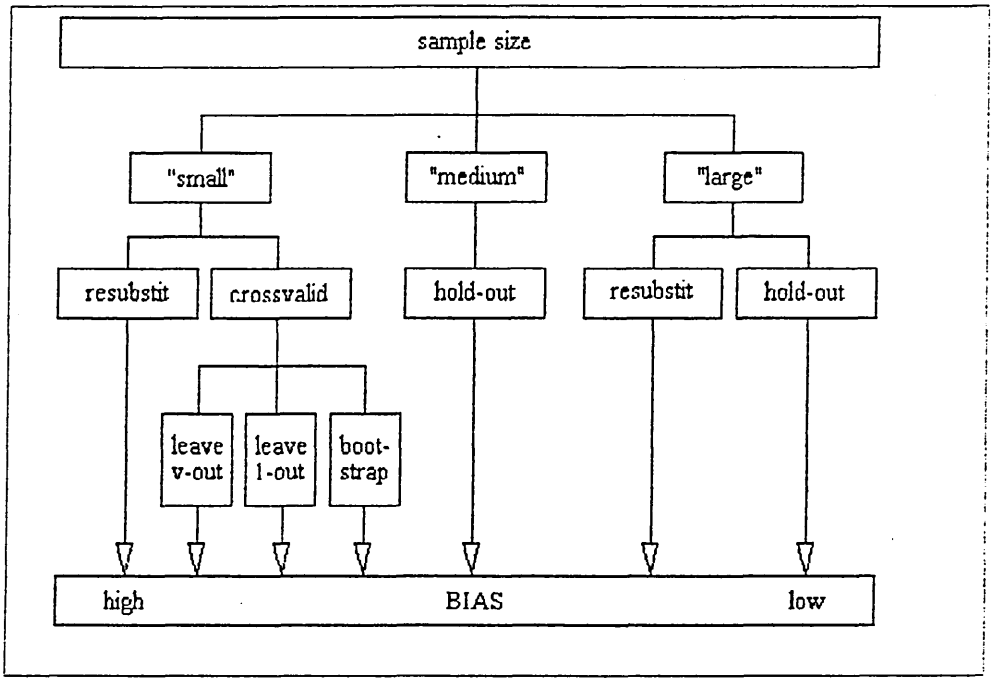


Figure 12.6-5: Crossvalidation decision tree

The above decision tree shows possible choices of crossvalidation techniques for different sample sizes. With smaller sample sizes more intricate crossvalidation techniques are required to achieve low absolute bias[45] in the estimated performance criteria. In figure 12.6-5 the terms "small", "medium" and "large" are deliberately put in quotes as it is not possible to quantify sample sizes in

[45] the sign of the bias will depend on which crossvalidation technique is used, e.g. resubstitution will lead to a negative bias on error rates as if it produces optimistic estimates.

absolute numbers. The consequences of various crossvalidation techniques on bias properties of performance criteria will ultimately also depend on structural characteristics of individual datasets. The above crossvalidation decision tree has been included in order to illustrate the general relationship between sample size and crossvalidation technique.

12.7 Conclusions

Five distinct factors determine choice of a discriminant procedure for discrete data: amount of data and number of groups, demands placed on the discriminant, cost and execution time constraints, model information and skill of the user applying a procedure. The demands placed on a discriminant play a central role in the selection process because the purposes for which a discriminant is wanted may differ. Characteristics such as low bias, high reliability, low variance or low error rates will frequently be given different priorities. These in turn will bear on selection. Assuming that skill and constraints factors may be neglected the initial starting point for procedure selection will be the theoretical model information and the technical data information. The selection "tree" includes a feedback loop that may be used after inspection of the first set of performance statistics. These may cover the whole range from $\varepsilon^{counting}$, $\varepsilon^{posterior}$ and η to the distribution $f(\tau)$ and the performance curves, $\varphi(\tau)$. Employment of suitable density estimation and crossvalidation techniques are used to obtain low bias and stable allocation rules and thus reliable performance statistics. Choice of an optimal discriminant procedure for discrete data is thus seen to consist of three different aspects: (1) initial choice largely on theoretical and technical grounds using a selection tree, (2) execution of the given discriminant procedure including density estimation and crossvalidation routines and (3) appraisal of the performance statistics. These three aspects are

shown symbolically as sides of a cube respectively in anticlockwise direction in figure 12.7-1.

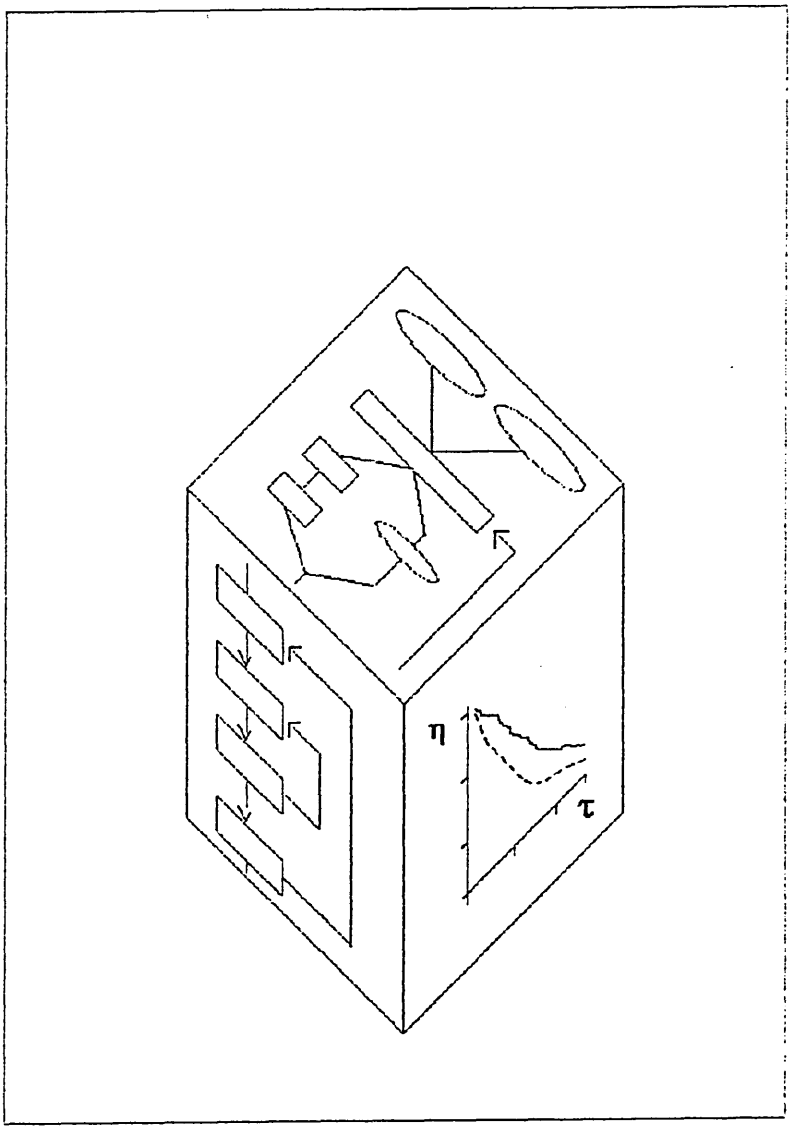


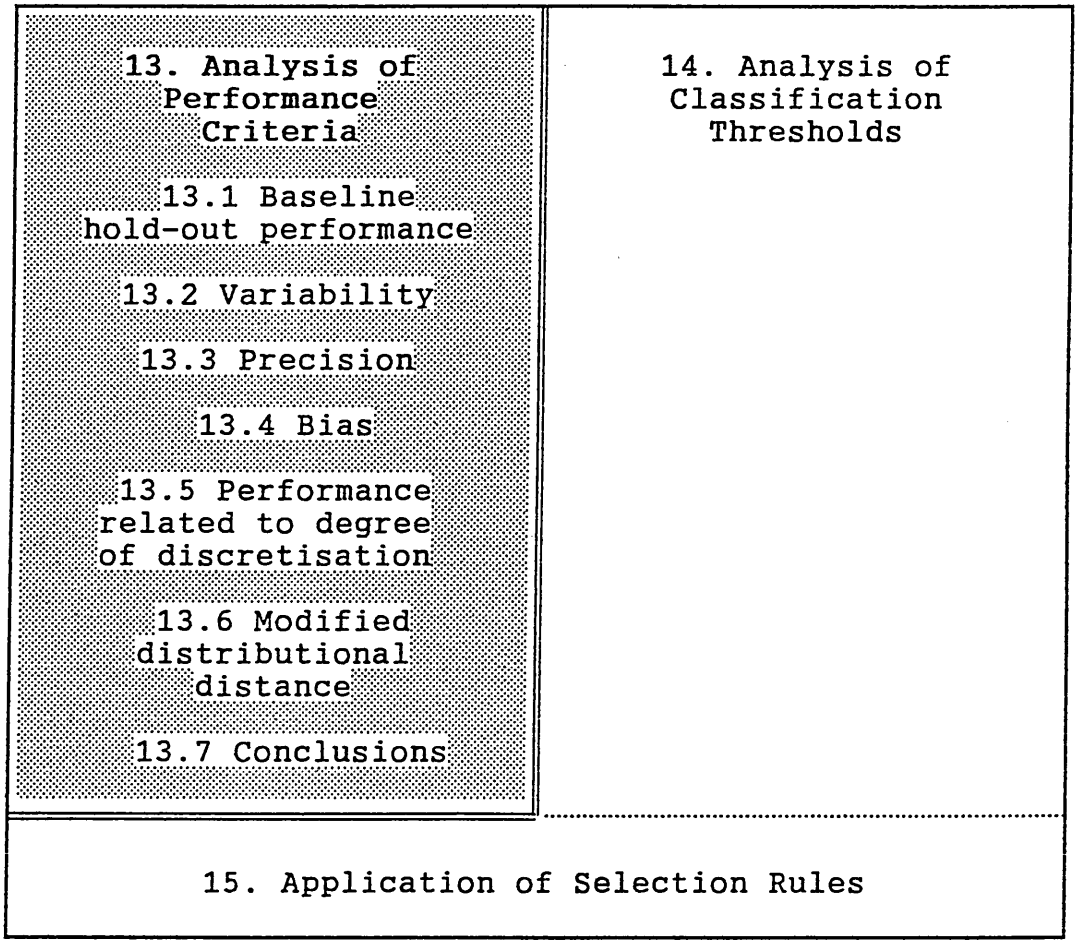
Figure 12.7-1: Aspects of procedure selection

I: INTRODUCTION

II: REVIEW

III: METHOD

IV: RESULTS



V: DISCUSSION

Results are reviewed in six sections. Section 13.1 gives baseline information on hold-out based levels of the three performance criteria $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η. Section 13.2 analyses variability of performance criteria averaged over datasets and discriminant procedure. Analyses of precision of the estimators are given in section 13.3. Analyses of bias appear in section 13.4. Section 13.5 gives results on performance as a function of *discreteness* of the dataset. Section 13.6 gives results for the modification[46] carried out on Goldstein and Dillon's distributional distance procedure.

Results for varying classification thresholds are treated in chapter 14 and consequences for procedure selection - though related to performance criteria - are dealt with separately in chapter 15. Comprehensive results are listed separately in the appendices A to G. Chapters 13 and 14 contain only selected extracts of the complete tables as well as summary statistics required for a general appraisal of the results.

13.1 Baseline hold-out performance

The tables in appendix A show comprehensive results for the misallocation errors, $\varepsilon_{counting}$ and $\varepsilon_{posterior}$ and the criterion, η, under *50* percent hold-out crossvalidation by dataset and discriminant procedure. The following tables 13.1-1, 13.1-2 and 13.1-3 are extracts of these tables showing baseline hold-out performance. They show estimates of performance criteria for all of the real datasets as well as for some of the artificial datasets. These datasets are listed in chapter 11 and are available in full detail from the author. The datasets are further divided into real

[46] See chapter 10, section 2.

and artificial data type and also classified by predominant level of predictor variable.

The artificial datasets with dichotomous predictors consist of the dataset given by Dillon and Goldstein (1978) for demonstrating the distributional distance procedure as well as the *MA4353* series of *10* samples generated by using the Bahadur expansion (chapters 4 and 10). The artificial data with polytomous predictors consist of the *BANANA* dataset designed to illustrate the need for curvilinear separation lines, the *INTERAC1* dataset designed to show the relevance of interactions for procedure selection and a sequence of datasets generated by discretising samples from the normal distribution (*NORMAL11 to NORMAL17*) designed to inspect the robustness of the linear discriminant function.

Discriminant procedures are grouped into direct and indirect classes. The hold-out crossvalidated performance criteria were calculated by averaging over *100* replications. This was done by dividing the original dataset into equal sized training and test datasets *100* times thus giving *100* estimates of hold-out performance. The Bahadur procedures were applied only to the dichotomous datasets due to technical admissibility. The logistic procedure was not calculated for the polytomous data response sets with *4* or more groups because it was programmed for analysis of data involving up to at most *3* populations (chapter 10). An extension is straightforward (chapter 4). The distributional distance procedure *DD2* with the modification suggested in chapter 10 was not applied to the sequence of artificial data with dichotomous predictors[47].

[47] Further results comparing the *DD1* and the *DD2* procedures are given in section 13.6.

expectation of hold-out based estimates of err(counting)			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lg1	nlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.340	0.341	0.340	0.343	0.341	0.354	0.344	0.353	0.344	0.343	0.360
		CESAR1	0.123	0.124	0.123	0.135	0.134	0.146	0.127	0.136	0.140	0.139	0.140
		GRADE	0.240	0.245	0.240	0.253	0.253	0.053	0.293	0.520	0.360	0.567	0.597
		LIZARD	0.111	0.113	0.114	0.126	0.111	0.140	0.113	0.141	0.109	-	0.169
		VIRGIN	0.200	0.207	0.209	0.204	0.209	0.220	0.210	0.205	0.204	0.297	0.204
	polytom.	CHD	-	-	-	0.069	0.069	-	0.069	0.340	0.270	0.322	0.510
		COLLEGE	-	-	-	0.201	0.200	-	0.203	0.274	0.215	0.215	0.234
		CREDIT	-	-	-	0.200	0.252	0.257	0.292	0.302	0.206	0.200	0.410
		EDUC	-	-	-	0.311	0.311	-	0.311	0.713	0.617	0.610	0.791
		ESTEEM	-	-	-	0.303	0.303	0.303	0.303	0.401	0.432	0.427	0.570
		IRIS	-	-	-	0.144	0.120	0.377	0.170	0.160	0.234	0.351	0.170
		KRETSCHM	-	-	-	0.360	0.309	0.355	0.424	0.366	0.204	0.205	0.414
		VOTING	-	-	-	0.167	0.172	-	0.169	0.104	0.160	0.309	0.103
artif.	dichot.	DILLON	0.004	0.005	0.005	0.006	0.090	0.200	0.091	0.496	0.094	0.420	0.366
		MA435300	0.440	0.443	0.443	0.370	0.420	0.504	0.369	0.451	0.301	-	0.366
		MA435301	0.435	0.441	0.441	0.360	0.416	0.540	0.346	0.439	0.293	-	0.345
		MA435302	0.399	0.400	0.403	0.303	0.407	0.472	0.374	0.399	0.313	-	0.373
		MA435303	0.402	0.400	0.399	0.370	0.377	0.460	0.356	0.400	0.302	-	0.363
		MA435304	0.390	0.405	0.401	0.346	0.366	0.443	0.339	0.405	0.200	-	0.341
		MA435305	0.432	0.427	0.420	0.357	0.419	0.497	0.350	0.427	0.202	-	0.346
		MA435306	0.400	0.403	0.401	0.399	0.401	0.450	0.377	0.401	0.290	-	0.375
		MA435307	0.433	0.425	0.432	0.359	0.393	0.495	0.352	0.430	0.294	-	0.349

Table 13.1-1: Estimates of err(counting)

expectation of hold-out based estimates of err(posterior_1)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dh1
type	pred	data											
real	dichot.	BREAST	0.283	0.282	0.282	0.348	0.322	0.322	0.332	0.331	0.338	0.338	0.333
		CESAR4	0.119	0.120	0.120	0.110	0.072	0.315	0.110	0.124	0.115	0.119	0.126
		GRADE	0.215	0.215	0.213	0.220	0.202	0.287	0.192	0.186	0.219	0.155	0.239
		LIZARD	0.110	0.109	0.106	0.127	0.103	0.110	0.109	0.119	0.094	-	0.106
		VIRGIN	0.198	0.200	0.196	0.223	0.144	0.230	0.201	0.222	0.100	0.210	0.236
	polytom.	CHD	-	-	-	0.069	0.069	-	0.069	0.068	0.071	0.069	0.068
		COLLEGE	-	-	-	0.207	0.106	-	0.197	0.198	0.199	0.200	0.192
		CREDIT	-	-	-	0.195	0.230	0.360	0.170	0.174	0.167	0.177	0.175
		EDUC	-	-	-	0.311	0.314	-	0.312	0.313	0.311	0.310	0.466
		ESTEEM	-	-	-	0.304	0.303	0.270	0.303	0.304	0.304	0.301	0.301
		IRIS	-	-	-	0.113	0.052	0.379	0.132	0.125	0.083	0.100	0.096
		KRETSCHM	-	-	-	0.230	0.194	0.411	0.275	0.134	0.163	0.193	0.242
		VOTING	-	-	-	0.100	0.150	-	0.169	0.171	0.175	0.179	0.177
artif.	dichot.	DILLON	0.091	0.089	0.092	0.082	0.087	0.153	0.076	0.073	0.084	0.093	0.081
		MA435300	0.327	0.327	0.321	0.332	0.371	0.365	0.200	0.304	0.260	-	0.266
		MA435301	0.267	0.256	0.250	0.309	0.359	0.376	0.265	0.225	0.203	-	0.240
		MA435302	0.222	0.210	0.210	0.320	0.336	0.307	0.200	0.206	0.296	-	0.290
		MA435303	0.250	0.251	0.252	0.306	0.321	0.309	0.200	0.220	0.274	-	0.200
		MA435304	0.266	0.266	0.266	0.310	0.331	0.370	0.270	0.276	0.277	-	0.254
		MA435305	0.269	0.270	0.270	0.313	0.349	0.376	0.255	0.250	0.239	-	0.200
		MA435306	0.232	0.233	0.236	0.315	0.336	0.406	0.200	0.265	0.202	-	0.265
		MA435307	0.202	0.294	0.205	0.316	0.357	0.365	0.263	0.201	0.269	-	0.249

Table 13.1-2: Estimates of err(posterior)

expectation of hold-out based estimates of eta			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.689	0.689	0.689	0.659	0.668	0.662	0.662	0.648	0.656	0.656	0.645
		CESAR1	0.879	0.878	0.879	0.877	0.897	0.769	0.877	0.865	0.868	0.868	0.868
		GRADE	0.774	0.775	0.775	0.770	0.778	0.493	0.765	0.518	0.656	0.481	0.479
		LIZARD	0.898	0.898	0.898	0.874	0.893	0.875	0.889	0.861	0.891	-	0.839
		VIRGIN	0.797	0.797	0.797	0.786	0.824	0.771	0.795	0.796	0.793	0.783	0.796
	polytom.	CHD	-	-	-	0.931	0.931	-	0.931	0.654	0.723	0.677	0.498
		COLLEGE	-	-	-	0.796	0.883	-	0.888	0.727	0.785	0.785	0.768
		CREDIT	-	-	-	0.763	0.755	0.687	0.765	0.689	0.765	0.772	0.649
		EDUC	-	-	-	0.718	0.718	-	0.709	0.358	0.425	0.426	0.298
		ESTEEM	-	-	-	0.697	0.697	0.718	0.697	0.519	0.569	0.573	0.435
		IRIS	-	-	-	0.872	0.918	0.665	0.852	0.829	0.756	0.639	0.853
		KRETSCHM	-	-	-	0.781	0.759	0.617	0.658	0.654	0.738	0.728	0.646
		VOTING	-	-	-	0.827	0.839	-	0.831	0.816	0.831	0.612	0.819
artif.	dichot.	DILLON	0.913	0.913	0.911	0.916	0.911	0.824	0.916	0.518	0.918	0.588	0.648
		MA435380	0.613	0.615	0.618	0.645	0.685	0.566	0.676	0.558	0.695	-	0.677
		MA435381	0.649	0.652	0.651	0.666	0.612	0.538	0.695	0.571	0.788	-	0.697
		MA435382	0.690	0.691	0.690	0.649	0.629	0.571	0.673	0.611	0.687	-	0.673
		MA435383	0.674	0.674	0.675	0.658	0.651	0.576	0.682	0.684	0.781	-	0.678
		MA435384	0.668	0.665	0.667	0.672	0.652	0.594	0.692	0.684	0.789	-	0.692
		MA435385	0.650	0.651	0.651	0.665	0.616	0.563	0.697	0.584	0.715	-	0.698
		MA435386	0.684	0.682	0.682	0.643	0.632	0.568	0.672	0.687	0.693	-	0.673
		MA435387	0.642	0.641	0.641	0.663	0.625	0.578	0.693	0.571	0.787	-	0.691

Table 13.1-3: Estimates of eta

Inspection of tables 13.1-1 to 13.1-3 reveals that the actual observed values of all performance criteria generally lie within the expected ranges: $0 \leq \varepsilon_{counting} \leq 1/2$, $0 \leq \varepsilon_{posterior} \leq 1/2$ and $1/2 \leq \eta \leq 1$. The performance criteria were deliberately constructed to have similar ranges thus making them more comparable with respect to their variances (chapter 10).

Deviations from the expected ranges are seen by the performance of the indirect distance based procedures for the *EDUC* dataset when assessed by $\varepsilon^{counting}$ and η (tables 13.1-1 and 13.1-3). Reasons for this considerable exception are assumed related to the fact that the *EDUC* dataset is characterised by an ordinal response variable and also by a wider range of prior probabilities ($\pi_1 = 0.192$, $\pi_2 = 0.066$, $\pi_3 = 0.054$, $\pi_4 = 0.689$). Similarly the performance of the first order logistic, *LG1*, is slightly worse than with chance allocation for the first two of the series of the artificially generated datasets *MA435300* and *MA435301* (table 13.1-1).

13.2 Variability of performance criteria

The tables in appendix B show comprehensive results for variability estimates of $\varepsilon^{counting}$, $\varepsilon^{posterior}$ and η for the resubstitution, hold-out and leave-one-out crossvalidation options. The following tables 13.2-1 and 13.2-2 present results under hold-out crossvalidation conditions extracted from the tables in the appendix. Variability is computed in terms of standard deviations of the respective performance criteria computed either across datasets or across direct discriminant procedures[48].

[48] The criterion η depends on posteriors computed via pseudo likelihoods derived from the same multinomial model (chapters 8 and 10). Standard deviations computed across indirect procedures for η would therefore lead to favourably underestimating variability of the eta criterion. To avoid this therefore indirect procedures are excluded from variability estimates.

var. of hold-out based performance averaged across direct procedures		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
type	data							
real	BREAST	1.4	0.005	0.3	0.026	2.1	0.014	7
	CESAR4	6.7	0.009	57.2	0.000	5.0	0.043	7
	CHD	0.4	0.000	0.3	0.000	0.0	0.000	3
	COLLEGE	1.7	0.003	5.3	0.011	0.5	0.004	3
	CREDIT	7.0	0.019	35.3	0.006	5.0	0.037	4
	EDUC	0.0	0.000	0.4	0.001	0.1	0.000	3
	ESTEEM	0.0	0.000	4.2	0.013	0.9	0.006	4
	GRADE	66.2	0.226	14.0	0.031	14.4	0.106	7
	IRIS	55.0	0.115	85.5	0.144	13.2	0.109	4
	KRETSCHM	13.2	0.040	33.6	0.094	9.1	0.062	4
	LIZARD	9.3	0.011	6.9	0.008	0.9	0.008	7
	VIRGIN	2.4	0.005	14.7	0.029	2.0	0.016	7
	VOTING	1.5	0.003	9.0	0.015	0.7	0.006	3
artif.	BANANA	80.0	0.097	72.1	0.079	9.9	0.000	4
	DILLON	41.4	0.043	27.0	0.026	3.8	0.034	7
	INTERAC1	85.8	0.215	82.4	0.151	21.8	0.171	4
	MA435300	10.7	0.046	9.1	0.030	5.5	0.034	7
	MA435301	15.5	0.066	16.9	0.050	7.9	0.050	7
	MA435302	7.7	0.031	23.9	0.068	6.8	0.045	7
	MA435303	8.3	0.033	17.4	0.051	5.6	0.037	7
	MA435304	9.5	0.037	13.7	0.041	4.7	0.031	7
	MA435305	12.0	0.050	15.6	0.047	6.6	0.042	7
	MA435306	6.1	0.025	22.5	0.065	6.5	0.042	7
	MA435307	12.0	0.049	12.6	0.039	5.9	0.037	7

Table 13.2-1: Variability of hold-out perf.

variability of hold-out based performance averaged across data sets		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
procedure class	discriminant							
direct	bh1	39.7	0.132	31.3	0.071	13.8	0.099	16
	bh2	39.6	0.131	31.6	0.072	13.8	0.099	16
	bh3	39.5	0.131	31.5	0.071	13.8	0.099	16
	ker	66.0	0.139	54.5	0.108	15.8	0.119	37
	ldf	64.9	0.147	68.9	0.133	17.5	0.139	37
	lg1	61.9	0.200	31.0	0.095	17.6	0.122	29
	mlt	59.5	0.133	55.5	0.098	14.2	0.114	37
indirect	cen	59.1	0.172	61.3	0.102	22.3	0.158	37
	dd1	68.4	0.139	61.1	0.102	16.3	0.129	36
	dd2	85.5	0.170	76.7	0.096	21.0	0.164	26
	dhl	65.1	0.181	62.6	0.109	21.7	0.162	37

Table 13.2-2: Variability of hold-out perf.

The number of datasets or respectively the number of procedures used in computing averages appears in the final (*count*) column of each table.

The expected range for η, $\{0.5,1.0\}$, is of the same width as the range for the misallocation errors yet the expected value of η is higher. Therefore its standard deviation will be expected to be correspondingly higher. For this reason the scale independent coefficients of variation, *cv*, are also included. Comparing these *cv*'s in tables 13.2-1 and 13.2-2 reveals that the variability of the suggested eta criterion is generally lower than that of the error rate estimators $\varepsilon^{counting}$ and $\varepsilon^{posterior}$. This finding reflects the symmetric construction of η (chapter 8). The standard deviations, or equivalently, the coefficients of variation for the two error rate estimators are of a similar order of magnitude. These results hold independently of type of crossvalidation technique used and also irrespective of whether averages are computed across direct procedures

(table 13.2-1) or across datasets (table 13.2-2). The above also applies for resubstitution and leaving-one-out based crossvalidation.

The generally lower variability of η, however, does not preclude its ability to show up differences in performance not detected by the customary error rate, $\varepsilon_{counting}$. This is illustrated in table 13.2-3 which has been extracted from tables 13.1-1 and 13.1-3 as well as the corresponding second parts of tables A-1 and A-3 from the appendix.

		Expectation of hold-out based estimates	
dataset	procedure	$\varepsilon_{counting}$	η
BREAST	*BH2*	*.341*	*.689*
	LDF	*.341*	*.668*
GRADE	*KER*	*.253*	*.770*
	LDF	*.253*	*.778*
LIZARD	*BH1*	*.111*	*.890*
	LDF	*.111*	*.893*
	LG1	*.140*	*.875*
	CEN	*.141*	*.861*
VIRGIN	*BH3*	*.209*	*.797*
	LDF	*.209*	*.824*
	KER	*.204*	*.786*
	DD1	*.204*	*.793*
ESTEEM	*LG1*	*.303*	*.710*
	MLT	*.303*	*.697*
MA435308	*KER*	*.326*	*.683*
	MLT	*.325*	*.714*
BANANA	*KER*	*.037*	*.959*
	DHL	*.037*	*.964*
NORMAL15	*KER*	*.052*	*.946*
	LDF	*.051*	*.957*

Table 13.2-3: Err(counting) compared to eta

Table 13.2-3 gives expected hold-out based performance for all cases where differences for $\varepsilon_{counting}$ are very small

but large by comparison for η. These results are to be expected from the hypothetical example of datasets *A*, *B* and *C* given in chapter 8.

13.3 Precision of performance criteria

The tables in appendix C show expected standard errors of conditional and unconditional estimates of the performance criteria for all combinations of dataset and discriminant procedure. In addition the estimates of standard errors have also been averaged over datasets, procedures and both as in section 13.2. A graphical comparison of unconditional estimates between the three performance criteria is given in figures 13.3-1, 13.3-2 and 13.3-3. In all these plots estimates of standard errors are based on *unconditional expectation*[49]. Standard errors have been computed from the series of $T = 100$ bootstrap trials. Each combination of direct discriminant procedure and real dataset is represented by a plus sign (+). The estimates are restricted to direct procedures for the same reasons as given in section 13.2. The estimates also refer only to *real* datasets in order to facilitate interpretation of the plots[50].

---

[49] For an explanation of *unconditional expectation* see section 13.4.

[50] The *artificial* data sets include several series of similar data sets, such as the *NORMAL11-NORMAL17* series. Exclusion eliminates artefacts due to the artificial nature of these data.

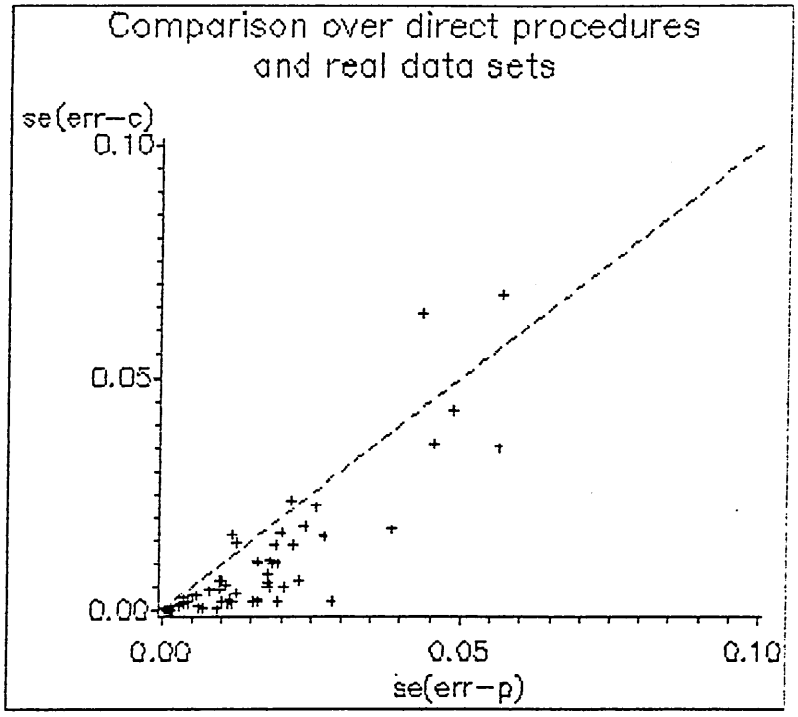


Figure 13.3-1: Standard error of err-c and err-p

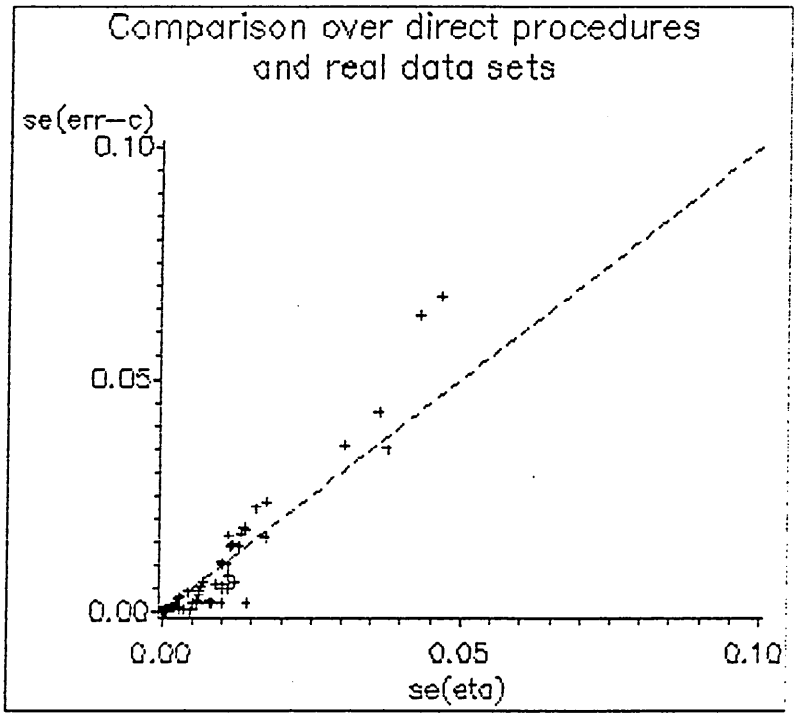


Figure 13.3-2: Standard error of err-c and eta

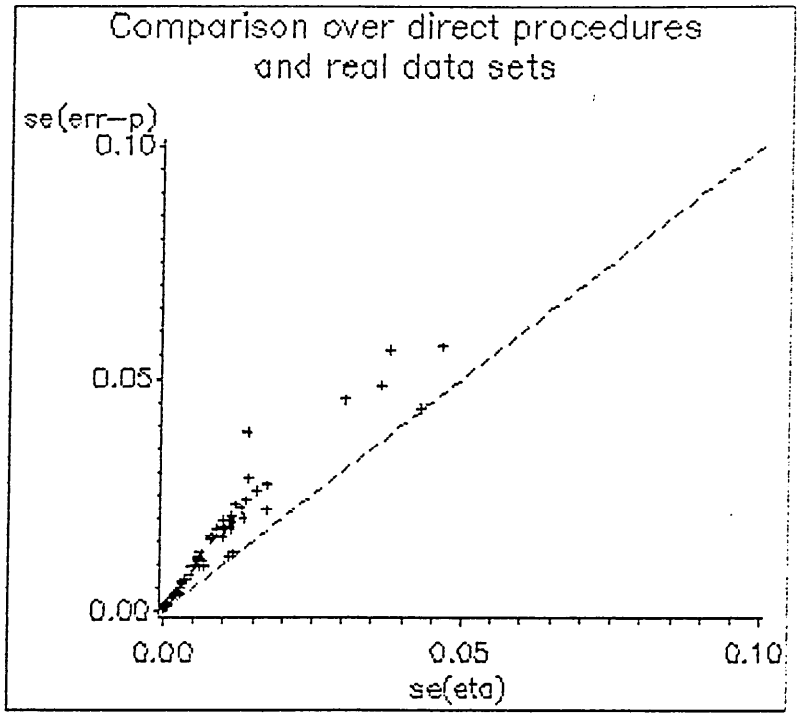


Figure 13.3-3: Standard error of err-p and eta

Figures 13.3-1 to 13.3-3 have uniformly scaled axes and a broken diagonal line drawn at *45* degrees for direct comparison. Inspection of the ranges for standard errors reveals that they are highest for $\varepsilon^{counting}$ (*0.068*), next highest for $\varepsilon^{posterior}$ (*0.056*) and smallest for η (*0.047*). Comparison between $se(\varepsilon^{counting})$ and $se(\varepsilon^{posterior})$ shows that the common counting based error rate exhibits lower expected standard errors than $\varepsilon^{posterior}$ as most observations lie beneath the diagonal line (see figure 13.3-1). This finding is surprising, in that it is contrary to what Hora and Wilcox (1982) suggested[51]. On these empirical grounds it must be therefore concluded that posterior probability based estimators need not always exhibit lower variance. The precision of $\varepsilon^{counting}$ in terms of standard errors is also better than that of η at the bottom end of the scale as well as in the upper range of standard errors beyond values of about *0.015* (see figure

[51] See also chapters 7 and 8 where the properties of posterior based performance criteria are discussed.

13.3-2). The superiority of η above $\varepsilon^{posterior}$ over the entire range is evident from figure 13.3-3.

In contrast to the variability discussion in 13.2 where averages were computed across datasets and procedures, the expected standard errors are derived by using the bootstrap method to obtain unconditional estimates. Estimates are thus averaged across bootstrap samples for each combination of dataset and procedures.

The interpretation of standard errors of performance criteria requires caution because a low standard error may mask differences in performance. Performance criteria with low standard errors may not distinguish where others with high standard errors do. On the other hand one would wish to have performance criteria that are sensitive to real differences yet reflect these with small bias and consistently, i.e. with small standard errors. As was pointed out in section 13.2 the empirical evidence reflected the designed properties of η. Thus interpreting the above low standard errors for η in conjunction with the ability of η to detect differences in performance where $\varepsilon^{counting}$ does not, is taken as evidence that the desirable properties expected of η in chapter 8 are empirically confirmed.

13.4 Bias of performance criteria

To assess the bias of $\varepsilon^{counting}$, $\varepsilon^{posterior}$ and η two variants of the bootstrap were used to obtain initial estimates of "true" values, $\varphi^{conditional}$ and $\varphi^{unconditional}$. *Conditional performance* (see chapter 7) was estimated by calculating the discriminant rule $\delta(\mathbf{x})$ once and testing it on a single trial of $R = 100$ bootstrap replications t^* derived from the training sample t. *Unconditional performance* was estimated by calculating $\delta^*(\mathbf{x})$ itself using $T = 100$ bootstrap trials and obtaining

for each one of these, estimates of performance criteria from $R = 100$ replicates. This follows the methodology described in chapter 10. The relative conditional biases are then estimated as

$$\text{bias}^{\text{conditional}}_{\text{hold-out}} = \frac{E\left[\varphi_{\text{hold-out}}\right] - \varphi^{\text{conditional}}}{\varphi^{\text{conditional}}} \qquad (13.4\text{-}1)$$

for hold-out crossvalidation where $\varphi_{\text{hold-out}}$[52] stands for hold-out based estimates. Corresponding relative hold-out based unconditional biases are similarly given by

$$\text{bias}^{\text{unconditional}}_{\text{hold-out}} = \frac{E\left[\varphi_{\text{hold-out}}\right] - \varphi^{\text{unconditional}}}{\varphi^{\text{unconditional}}}. \qquad (13.4\text{-}2)$$

The analysis of bias characteristics of the performance criteria was deliberately conducted under hold-out crossvalidation conditions because this is a common crossvalidation technique (see section 1 of chapter 7). In expressions 13.4-1 and 13.4-2 above the values for $\varphi^{\text{hold-out}}$ correspond to the results presented in section 13.1. Both biases are expressed as percentage deviations in the following tables. Tables 13.4-1 to 13.4-4 are again extracts from more comprehensive tables listed in appendices D and E[53]. Table 13.4-1 shows averages of conditional bias estimates computed across direct procedures[54] for real and artificial datasets.

---

[52] The greek letter φ designates performance criteria in general.

[53] Appendix D gives results for bias estimates based on the conditional performance of the discriminant procedure while appendix E refers to unconditional performance.

[54] The comparison is again restricted to the class of direct procedures for the same reason as given in section 13.1.

bias of hold-out based conditional performance averaged across direct			err_c	err_p	eta	count
			%	%	%	n
type	pred	data				
real	dichot.	BREAST	0.892	0.056	-0.236	7
		CESAR4	5.300	-0.542	-0.384	7
		GRADE	19.017	4.249	-2.082	7
		LIZARD	4.425	1.680	-0.363	7
		VIRGIN	2.576	-1.119	-0.183	7
	polytom.	CHD	-0.241	-0.290	0.018	3
		COLLEGE	0.936	-0.345	-0.075	3
		CREDIT	28.041	0.431	-3.728	4
		EDUC	-0.011	0.117	-0.014	3
		ESTEEM	-0.008	0.366	-0.070	4
		IRIS	34.381	11.178	-2.828	4
		KRETSCHM	76.818	11.195	-11.656	4
		VOTING	0.634	0.271	-0.096	3
artif.	dichot.	DILLON	-0.415	0.122	0.499	7
		MA435300	13.666	2.062	-3.899	7
		MA435301	6.224	0.000	-1.611	7
		MA435302	5.085	-0.790	-0.252	7
		MA435303	3.724	-2.109	0.946	7
		MA435304	5.004	0.042	-0.998	7
		MA435305	8.344	-1.571	-1.526	7
		MA435306	8.437	-3.717	-0.752	7
		MA435307	5.623	-1.170	-0.729	7
		MA435308	5.942	-0.001	-1.186	7
		MA435309	7.141	-3.215	-0.469	7

Table 13.4-1: Relative bias of perf. criteria

bias of hold-out based unconditional performance averaged across direct			err_c	err_p	eta	count
			%	%	%	n
type	pred	data				
real	dichot.	BREAST	0.760	0.295	-0.257	7
		CESAR4	-0.804	0.299	0.076	7
		GRADE	13.577	5.981	-2.650	7
		LIZARD	1.839	1.782	-0.215	7
		VIRGIN	1.938	0.296	-0.314	7
	polytom.	CHD	-0.241	-0.528	0.025	3
		COLLEGE	0.780	-0.099	-0.083	3
		CREDIT	16.824	5.339	-3.297	4
		EDUC	-0.021	0.117	-0.005	3
		ESTEEM	-0.008	-0.070	0.014	4
		IRIS	18.682	18.732	-2.842	4
		KRETSCHM	42.661	18.892	-9.753	4
		VOTING	0.195	0.376	-0.055	3
artif.	dichot.	DILLON	3.551	1.705	-0.338	7
		MA435300	7.174	3.480	-2.792	7
		MA435301	6.223	1.686	-2.089	7
		MA435302	4.982	0.738	-1.410	7
		MA435303	6.243	1.552	-1.987	7
		MA435304	4.463	0.806	-1.231	7
		MA435305	6.325	1.215	-1.903	7
		MA435306	6.304	-0.242	-1.551	7
		MA435307	5.024	1.873	-1.613	7
		MA435308	6.755	0.913	-2.030	7
		MA435309	6.988	0.730	-1.770	7

Table 13.4-2: Relative bias of perf. criteria

bias of hold-out based conditional performance averaged across data sets		err_c	err_p	eta	count
		%	%	%	n
procedure class	discriminant				
direct	bh1	2.813	0.150	-0.599	16
	bh2	1.929	0.065	-0.437	16
	bh3	2.280	-0.079	-0.475	16
	ker	36.076	127.396	-2.973	37
	ldf	4.047	6.666	-0.003	37
	lg1	18.663	15.480	-1.951	29
	mlt	104.784	77.155	-3.503	37
indirect	cen	138.449	25.996	-14.883	37
	dd1	19.283	14.925	-4.751	36
	dd2	74.298	24.738	-11.369	26
	dhl	150.521	34.402	-10.085	37

Table 13.4-3: Relative bias of perf. criteria

bias of hold-out based unconditional performance averaged across data sets		err_c	err_p	eta	count
		%	%	%	n
procedure class	discriminant				
direct	bh1	2.011	0.755	-0.594	16
	bh2	1.935	0.611	-0.518	16
	bh3	2.353	-0.553	-0.343	16
	ker	19.094	22.968	-2.508	37
	ldf	4.751	-0.186	-0.715	37
	lg1	2.360	14.711	-2.691	29
	mlt	29.387	10.563	-3.082	37
indirect	cen	89.588	2.437	-14.443	37
	dd1	6.295	-0.986	-4.314	36
	dd2	55.244	-4.147	-10.512	26
	dhl	69.894	8.173	-9.682	37

Table 13.4-4: Relative bias of perf. criteria

The number of datasets or respectively the number of procedures used in computing averages appears in the final (*count*) column of each table. All tables give bias estimates for the three performance criteria in adjacent columns. Inspection of tables 13.4-1 to 13.4-4 reveals that the hold-out based bias estimates $bias^{hold\text{-}out}$ generally show positive sign for the error rate estimates $\varepsilon^{counting}$ and $\varepsilon^{posterior}$ but negative sign for η. This implies that error rates are overestimated while the eta criterion is underestimated which is exactly in line with what is to be expected on theoretical grounds (see chapter 8). As was found in chapter 7 the expected hold-out based *actual* or *conditional* error is asymptotically at least as large as the *optimum* or *true* error. Table 13.4-2 shows corresponding results for $bias^{hold\text{-}out}$ based on unconditional estimates of the performance criteria φ. Note that the difference in bias estimates between conditional and unconditional expectation (tables 13.4-1 and 13.4-2) is more marked for the error rate estimates than for η. Overall η shows the smallest absolute bias. The results for the absolute values of bias estimates averaged across direct procedures from tables 13.4-1 and 13.4-2 may be summarised as shown in table 13.4-5.

type of data	conditional	unconditional
real	$b(\varepsilon^c) > b(\varepsilon^p) \cong b(\eta)$	$b(\varepsilon^c) \ncong b(\varepsilon^p) > b(\eta)$
artificial art(*NORMAL*)[55]	$b(\varepsilon^c) \cong b(\varepsilon^p) > b(\eta)$	$b(\varepsilon^c) > b(\varepsilon^p) \cong b(\eta)$ $b(\varepsilon^c) \cong b(\varepsilon^p) > b(\eta)$

Table 13.4-5: Absolute bias by dataset

In table 13.4-5 $b(\varepsilon^c)$ is shorthand for the absolute size of $bias(\varepsilon^{counting})$ and $\cong$ stands for *of approximately equal order of magnitude* while $\ncong$ means *neither is consistently larger than the other.* The other symbols are to be interpreted in analogous fashion. Inspection of table

[55] The artificial data series *NORMAL11* to *NORMAL17*

13.4-5 shows that throughout $b(\eta)$ is either lowest or at most of the same order of magnitude as $b(\varepsilon^{posterior})$.

A similar picture emerges from the corresponding estimates of $bias^{hold\text{-}out}$ averaged over datasets (tables 13.4-3 and 13.4-4). Again the absolute size of bias for η is generally small. This holds particularly, with the exception of the Bahadur and kernel based procedures, when estimates are based on conditional performance (table 13.4-3) and $b(\varepsilon^c) > b(\varepsilon^p) > b(\eta)$. In the case of unconditional performance (table 13.4-4) the bias estimates for the posterior based error estimator and for the eta criterion are of a similar order of magnitude: $b(\varepsilon^p) \cong b(\eta)$. These relationships have been summarised in a similar fashion in table 13.4-6.

procedure	conditional	unconditional
direct[56]	$b(\varepsilon^c) > b(\varepsilon^p) > b(\eta)$	-
BAHADUR	$b(\varepsilon^c) > b(\eta) > b(\varepsilon^p)$	-
KERNEL	$b(\varepsilon^p) > b(\varepsilon^c) > b(\eta)$	-
direct[57]	-	$b(\varepsilon^c) > b(\varepsilon^p) > b(\eta)$
LDF	-	$b(\varepsilon^c) > b(\eta) > b(\varepsilon^p)$
LG1	-	$b(\varepsilon^p) > b(\varepsilon^c) \cong b(\eta)$
indirect	$b(\varepsilon^c) > b(\varepsilon^p) > b(\eta)$	$b(\varepsilon^c) > b(\eta) > b(\varepsilon^p)$

Table 13.4-6: Absolute bias of procedures

The emerging pattern is not quite as clear as for the conditional bias estimates, yet in summary it is evident that with the sole exception of the logistic procedure the absolute bias of η is always smaller than the absolute bias of $\varepsilon^{counting}$. It is further at a minimum for all the direct procedures apart from the linear discriminant and the logistic.

---

[56] excluding the Bahadur and kernel based procedures

[57] excluding the linear discriminant and logistic procedures

13.5 Performance related to degree of discretisation

In order to inspect the behaviour of performance criteria for different discriminant procedures in relation to departures from normality in discrete datasets a specially designed series of artificial datasets (*NORMAL11* to *NORMAL17*) was constructed (chapter 11). Starting from the same continuous normal bivariate sample in each case the data are *discretised* by dividing the original continuous distribution for each independent variable X_j into $m \geq 2$ equidistant intervals and relabelling these with ordinal numbers. When the number m is large the modification in the resultant distribution is small. When m is small the modification is large. The degree of discreteness increases as m approaches *2*.

The dataset with highest index (*NORMAL17*) has highest discretisation level with *s=4* states while the dataset with the lowest index (*NORMAL11*) has *s=96* states and thus approximates closest to the - originally - normal distribution. Further details on these datasets are given in chapter 11. Figures 13.5-1, 13.5-2 and 13.5-3 show comparisons between five direct and three indirect discriminant procedures for estimates of $\varepsilon^{counting}$, $\varepsilon^{posterior}$ and η based on hold-out crossvalidation. Expected values are plotted against discretisation level with *1 = low* and *7 = high*.

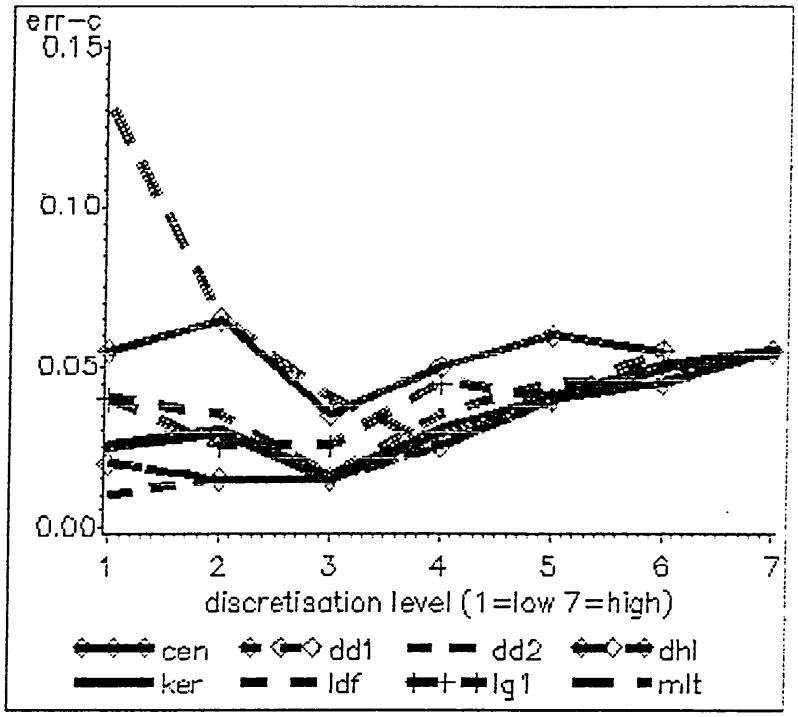


Figure 13.5-1: Hold-out err(counting)

The plot for $\varepsilon_{counting}$ in figure 13.5-1 shows that at low discretisation levels (when *m* is small and the discrete datasets have many states) the spread among estimates for different procedures is large. As the number of states decreases with increasing discretisation level all estimates tend towards the same value. Note also that the error rates for the linear discriminant rise steadily as the data structure progressively departs from normality. A high degree of differentiation in the data, as given at low discretisation levels, does not necessarily imply good performance for all procedures. This is very clear for the *MLT* procedure where performance improves initially towards level *4* and then deteriorates again towards level *7*. This is interpreted as due to the large number of cells, and thus parameters, requiring estimation for the multinomial.

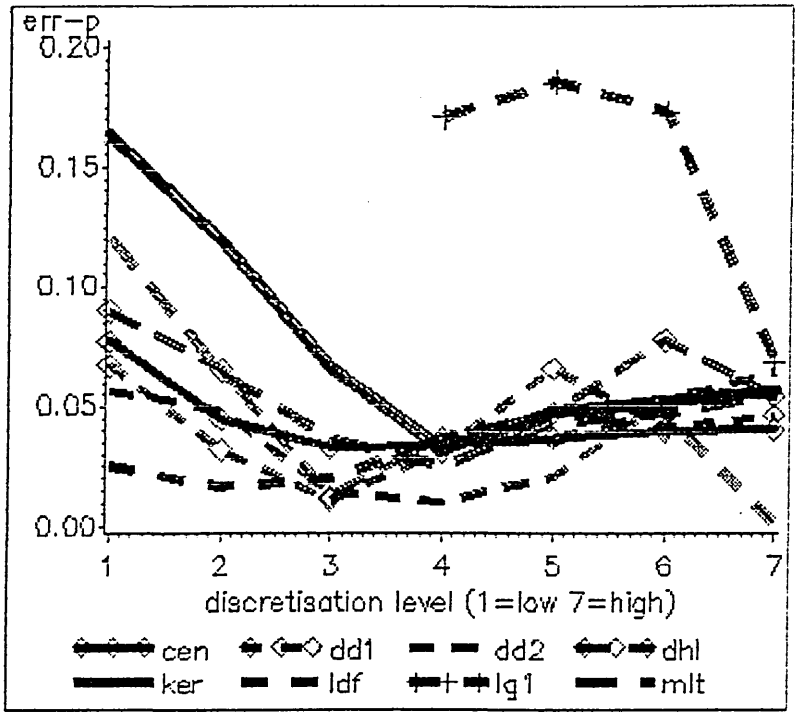


Figure 13.5-2: Hold-out err(posterior)

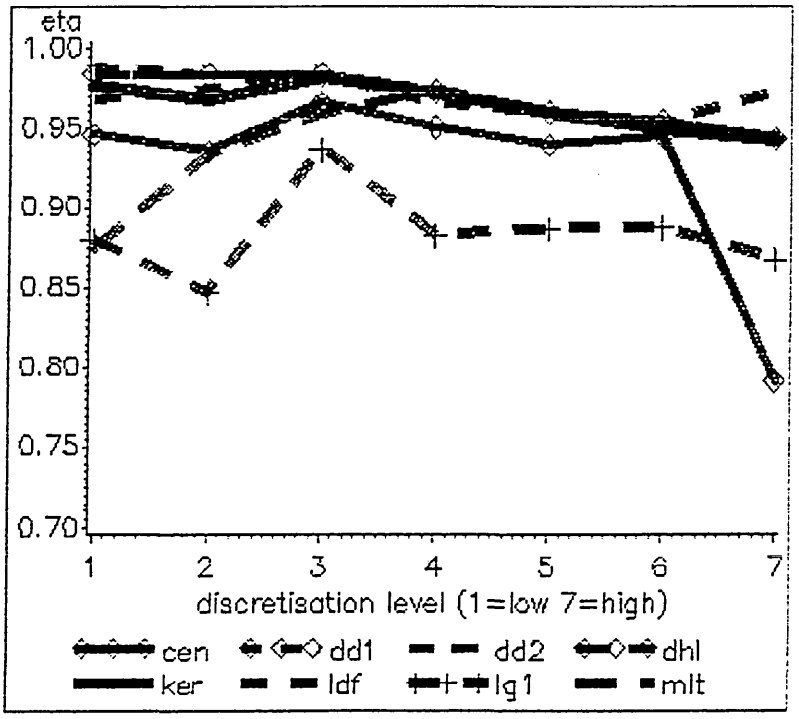


Figure 13.5-3: Hold-out η

Inspection of figures 13.5-2 and 13.5-3 reveals a similar pattern, although not quite as pronounced as for $\varepsilon^{counting}$.

The above findings for the hold-out based estimates also hold generally under leave-*1*-out crossvalidation conditions as seen from appendix F. On theoretical grounds it is to be expected that initially (at low discretisation levels) the linear discriminant function based discriminant will perform satisfactorily due to its robustness (chapter 4) and also because these data are here still closest to normal. On the other hand the centroid and the first order logistic discriminant procedures perform poorly on all of these artificial datasets. The respective performance of different discriminant procedures for the *NORMAL11* to *NORMAL17* series comes out more clearly for the performance criteria based on posterior probabilities, $\varepsilon^{posterior}$ and η (figures 13.5-2 and 13.5-3). Here the poorer performance of the logistic procedure is clearly indicated[58]. The linear discriminant performs particularly well across all discretisation levels when assessed by η. The dashed line for the *LDF* is generally highest in figure 13.5-3. It is especially noticeable that the eta criterion differentiates best between the eight discriminant procedures which is most evident at the higher discretisation levels.

The ability of the η criterion to "pick out" the linear discriminant procedure in applications to discretised originally continuous normal datasets is confirmed again by the results for the artificial series *NORMAL01*, *NORMAL02* and *NORMAL03*. Basic characteristics for these datasets are summarised from chapter 11 in table 13.5-1.

dataset	*N*	*s*	*q*
NORMAL01	*200*	*76*	*3*
NORMAL02	*80*	*19*	*3*
NORMAL03	*2000*	*137*	*2*

Table 13.5-1: *NORMAL01*, *NORMAL02*, *NORMAL03* data

[58] The vertical scale has been deliberately scaled to *0.00* to *0.20* thus causing only part of the estimates for the logistic to be shown.

From appendix A expected hold-out based performance estimates were used to compile table 13.5-2 showing the ranking of discriminant procedures. Procedures yielding optimal performance are listed first. The ranked lists stop as soon as the linear discriminant is reached.

*d*ataset	$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
NORMAL01	*DD2* *DD1* *CEN* *LDF*	*DD2* *DD1* *KER* *DHL* *LDF*	*LDF*
NORMAL02	*DD1* *DD2* *LDF*	*DHL* *DD1* *DD2* *MLT* *CEN* *LDF*	*DD1* *DD2* *LDF*
NORMAL03	*DD1* *DD2* *LDF*	*LDF*	*LDF*

Table 13.5-2: Ranked hold-out performance

The above table shows that the linear discriminant is singled out fairly rapidly by the η criterion in terms of expected hold-out based conditional performance. Inspection of the appropriate tables from the appendix for estimates of conditional performance similarly shows that in terms of standard error and bias again the linear discriminant procedure is consistently favoured by the η criterion. The tables also show that the linear discriminant also performs well in terms of $\varepsilon_{posterior}$.

13.6 Modified distributional distance

The modification of the original *distributional distance* procedure (*DD1*) of Goldstein and Dillon (1978) leading to the *DD2* procedure is described in chapter 10. By weighting the distance function

$$d_j^{DD1} = \left(\sqrt{p_{1j}} - \sqrt{p_{2j}}\right)^2 \qquad (13.6\text{-}1)$$

with the average of the cell proportions p_{ij} it was hoped to achieve greater stability in the performance of the *DD2* procedure.

$$d_j^{DD2} = \left(\frac{p_{1j}+p_{2j}}{2}\right)\left(\sqrt{p_{1j}} - \sqrt{p_{2j}}\right)^2 \qquad (13.6\text{-}2)$$

To test this hypothesis estimates of unconditional performance were compared for the *DD1* and the modified *DD2* procedure using equations 13.6-1 and 13.6-2 respectively. In the following plots estimates of bias and standard error of $\varepsilon^{counting}$ are shown for all datasets plotted against number of discrete states, s[59]. It is expected that as s increases there will be an increasing number of cells with smaller cell probabilities. In such cases the *DD2* procedure is expected to exhibit better performance. The first two figures 13.6-1 and 13.6-2 show expected bias and standard error of $\varepsilon^{counting}$ against number of discrete states, s.

[59] Two data sets with more than 100 states were excluded in order to blow up the scale at the lower end. Exclusion does not essentially distort the overall findings.

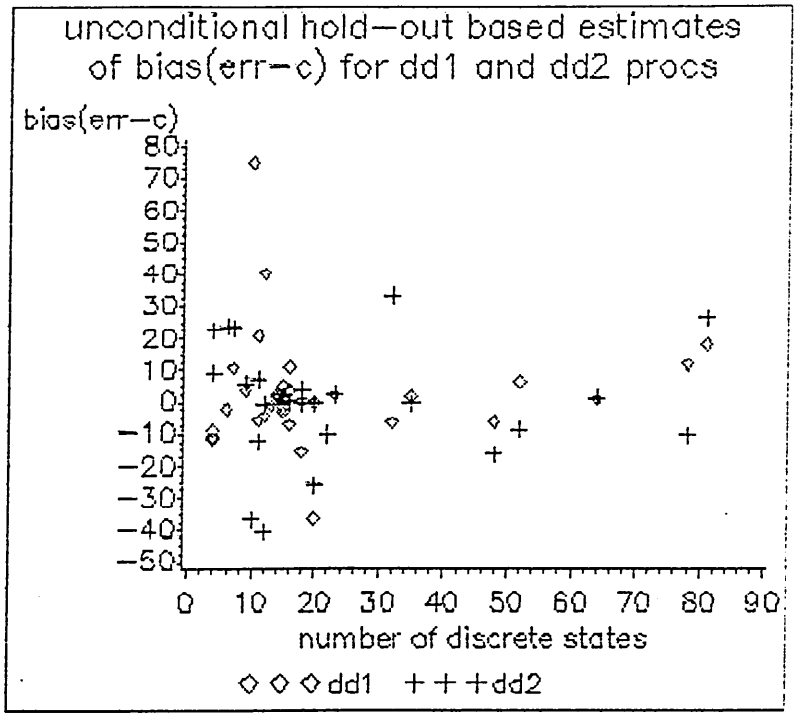


Figure 13.6-1: bias for *DD1* and *DD2* procs

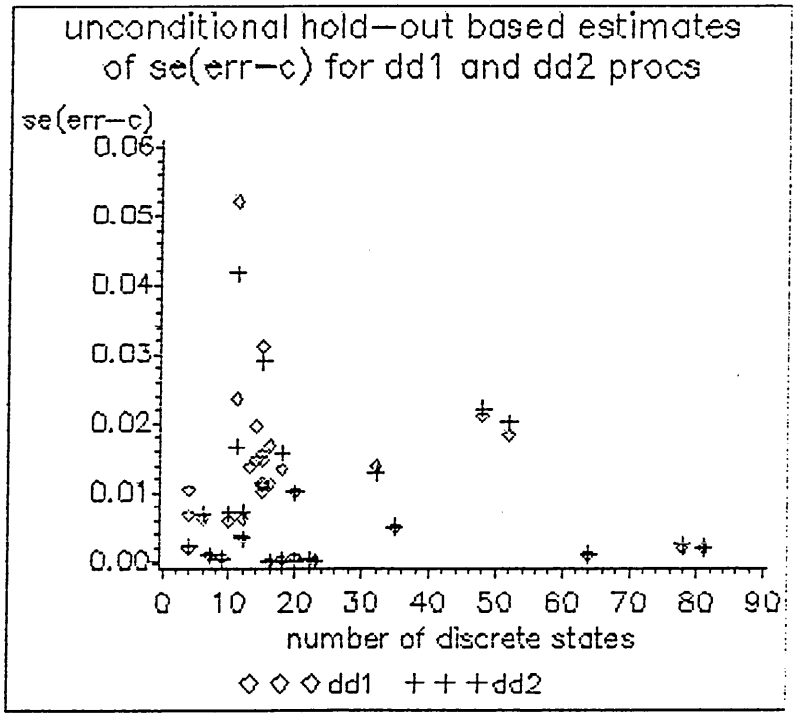


Figure 13.6-2: se for *DD1* and *DD2* procs

From both plots a clear pattern of decreasing absolute bias and standard error with increasing *s* emerges. Thus, generally, performance for both procedures *DD1* and *DD2* increases with number of discrete states. Differences in performance between the two indirect distance based procedures however are not readily apparent. To make these more clear the following plots in figures 13.6-3 and 13.6-4 were constructed to show absolute differences in estimates of bias and standard error for $\varepsilon^{counting}$ against number of states, *s*. The *delta(bias)* values plotted vertically are $\Delta_{DD1,DD2} = \hat{\varphi}_{DD1} - \hat{\varphi}_{DD2}$, where $\hat{\varphi}$ stands for *expected performance*[60]. Negative values of $\Delta_{DD1,DD2}$ thus indicate better performance for the *DD1* procedure.

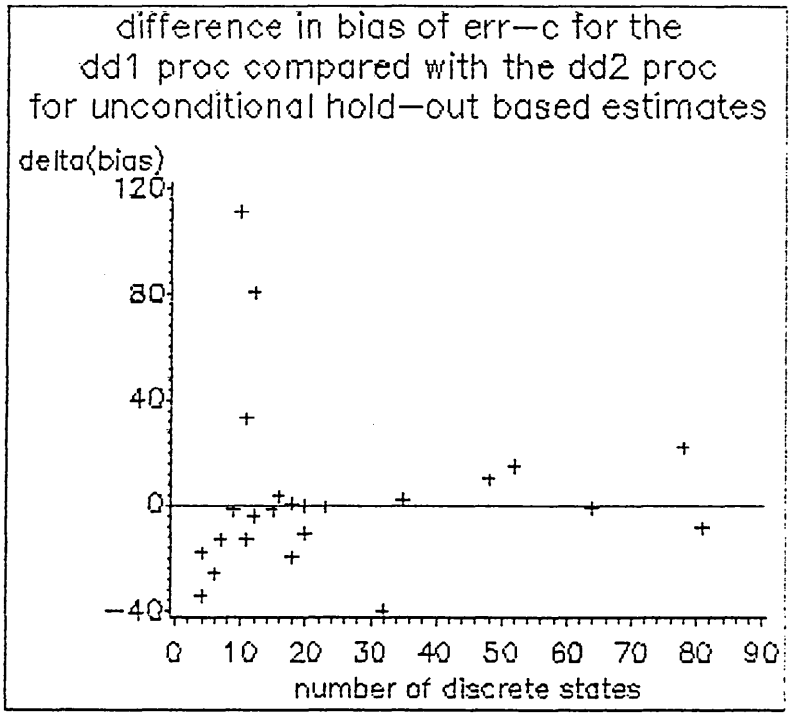


Figure 13.6-3: Delta(bias) for *DD1* and *DD2* procs

[60] In this case either bias or standard error.

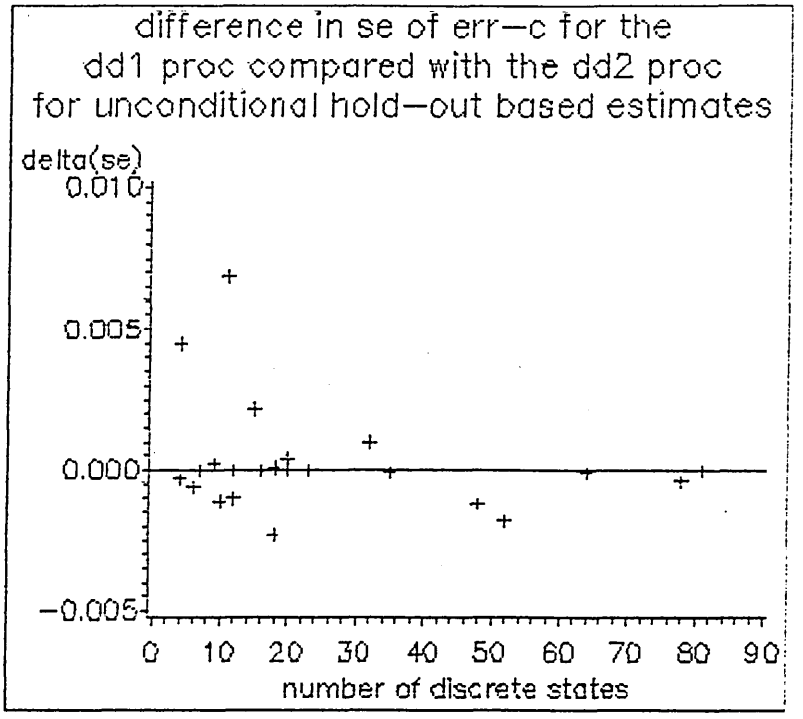


Figure 13.6-4: Delta(se) for *DD1* and *DD2* procs

Figure 13.6-3 shows that apart from three distinct exceptions the original *DD1* procedure has better bias characteristics for datasets with small values of *s*. At higher values of *s* the *DD2* procedure tends to be slightly better. With respect to the standard error the differences between procedures are not quite as clear. With the exception of three datasets with low values for *s* there is little difference between the *DD1* procedure and the modified *DD2* version in terms of standard errors.

13.7 Conclusions

Hold-out based performance lies generally within the expected ranges for $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η. The variability of performance criteria across direct procedures and datasets shows comparable coefficients of variation, *cv*, for $\varepsilon_{counting}$ and $\varepsilon_{posterior}$ whereas the *cv*'s for η are generally lower. This is seen to be mainly

the consequence of the symmetrical way in which η is constructed[61].

The fact that this does not, however, imply lower sensitivity of η as was seen in section 13.2 may be interpreted as empirical evidence of the expected characteristics of this posterior probability based performance criterion.

The results for the standard errors, $se(\varphi u)$, of unconditional estimates of performance criteria are not quite so clear. The $se(\varepsilon^{counting})$ is only rarely lower than $se(\varepsilon^{posterior})$ which appears to be contrary to the expectations of Hora and Wilcox (1982). Direct comparability is however not possible for four reasons: Hora and Wilcox based their findings entirely on Monte Carlo simulation studies of artificial data, they exclusively used discrimination between $g = 3$ populations, they used continuous data and based their findings largely on comparisons with the apparent - or resubstitution - error. The $se(\eta)$ is higher than $se(\varepsilon^{counting})$ at low values, yet lower at higher values. The absolute bias of η is generally smaller than the bias of $\varepsilon^{counting}$, irrespective of whether estimates are based on conditional or unconditional performance. The bias of η is also generally at most of a similar order of magnitude as the bias of $\varepsilon^{posterior}$.

Inspection of the behaviour of the performance criteria $\varepsilon^{counting}$, $\varepsilon^{posterior}$ and η for different degrees of discretisation - here indicating departures from normality - confirms the well known robustness of the linear discriminant function. Compared with the other two criteria η differentiates best between procedures across all discretisation levels.

---

[61] see chapter 8

Although the estimated misallocation errors, $\varepsilon^{counting}$, are generally smallest for the distance procedures for the discretised series of datasets (*NORMAL01*, *NORMAL02* and *NORMAL03*) it is found that especially the posterior based η criterion consistently singles out the linear discriminant procedure as optimal in terms of expected value, standard error and bias for these originally normally distributed continuous datasets. The posterior based error rate estimator, $\varepsilon^{posterior}$, also generally points to the linear discriminant procedure. This feature is attributed directly to the fact that these two performance criteria are computed from the posterior probabilities. The fact that the η criterion is even more consistent in this respect is again seen to be a consequence that η is based on the *entire distribution of posteriors* across discrete states as opposed to $\varepsilon^{posterior}$ which only depends on the posteriors for correctly allocated objects.

The expectations of the modifications of the distributional distance procedure of Goldstein and Dillon (1978), *DD1*, leading to the *DD2* procedure were not convincingly borne out by the results. It was expected that the *DD2* procedure would perform better for datasets with larger number of discrete states, s. The fact that this did not occur is assumed to be a consequence of the cumulative nature of the performance criteria. For the η criterion, for instance, the relative weights of individual states enter via the conditional densities, $f_i(\mathbf{x})$, as may be seen from the respective definition which is reproduced below from chapter 8 for reference:

$$\eta_i = \int \xi_i(\mathbf{x})\ f(\Pi_i|\mathbf{x})\ f_i(\mathbf{x})\ d\mathbf{x}\ . \qquad (\ 13.7\text{-}1)$$

The final η criterion is then a sum of individual η_i contributions weighted by the prior probabilities, π_i. Similar expressions (see chapter 8) ensure that the contributions to individual states are appropriately weighted by the joint densities.

I: INTRODUCTION

II: REVIEW

III: METHOD

IV: RESULTS

13. Analysis of Performance Criteria

14. Analysis of Classification Thresholds

14.1 Selection using classification thresholds

14.2 Selection for the *CESAR4* data

14.3 Selection for the *NORMAL16* data

14.4 Selection for the *GRADE* data

14.5 Selection for the *IRIS* data

14.6 Conclusions

15. Application of Selection Rules

V: DISCUSSION

Chapter 14 - Use of classification thresholds

This chapter is divided into six sections. Section 14.1 discusses the plots given in appendix G and outlines how choice of discriminant procedure can be aided using classification thresholds on the basis of the ideas presented in chapter 9. As classification thresholds are computed from posterior probabilities this technique can only be applied to direct procedures[62]. Sections 14.2 and 14.3 demonstrate the concept for a real dataset, *CESAR4*, and for an artificial dataset, *NORMAL16*, respectively. Both these datasets have $g = 2$ populations. Performance characteristics for the real *GRADE* and *IRIS* datasets are included in sections 14.4 and 14.5 because these datasets relate to discrimination between $g \geq 3$ populations where the variable classification threshold is expected to be more suitable (see chapter 9). For all datasets the focus will be particularly on the behaviour of the performance criteria at low relative thresholds in order to keep performance at acceptable levels. The comprehensive detailed results referred to are all listed in appendix G. For clarity the major plots discussed in the text are repeated below. Section 14.6 summarises conclusions that may be drawn from use of classification threshold analysis particularly in comparison with the non-thresholded performance criteria reported in chapter 13 and applications of the selection in chapter 15.

14.1 Selection using classification thresholds

Looking through the plots in appendix G perhaps the first thing to notice is the reciprocal nature of plots for the errors $\varepsilon_{counting}$ and $\varepsilon_{posterior}$ and the η criterion. This is however inherent in the definition of these criteria

---

[62] The lines referring to the indirect procedures DD1, DD2, DHL and CEN are the result of computing τ from the pseudo posteriors (chapter 10) and are not discussed at this stage.

(chapter 8) which means that high values of $\varepsilon^{counting}$ and $\varepsilon^{posterior}$ correspond to low values of η and vice versa. Note that the leave-one-out crossvalidation technique tends to give smoother distributions than the other less computing intensive techniques. This is marked especially for the *NORMAL16* dataset where the curves are more easily distinguished from one another. Comparing plots for different crossvalidation techniques for the same datasets it is apparent that beyond that there is little difference between crossvalidation techniques. The curves for the three performance criteria are also highly correlated.

The plots included in appendix G come in four series, one for each of the *CESAR4*, *NORMAL16*, *GRADE* and *IRIS* datasets. Each series begins with three plots of the estimated distribution of relative posterior differences $\hat{f}(\tau)$. The first of these plots shows the distribution across the entire range while the two following ones respectively enlarge selected regions from the first plot in order to highlight differences between procedures. A reliable and thus effective discriminant procedure will exhibit $\hat{f}(\tau)$ distributions with a predominance of large relative differences, τ. Thus $\hat{f}(\tau)$ distributions with positive skew will point to procedures with good discriminatory ability. The distributions of $\hat{f}(\tau)$ are derived using the bootstrap techniques outlined in chapter 9.

Procedure selection is carried out in two stages. First the estimated distributions of $\hat{f}(\tau)$ are inspected for *degree of positive skew* thus allowing initial procedure choice. Next the individual performance curves are judged for both crossvalidation techniques. *Absolute level and rate of change* of performance criteria in the first half of the τ range are used as further guidelines to narrow down choice. The results of these inspections are summarised in tabular form at the end of sections 14.2 to 14.5 to enable comparison. From these tables conclusions may be drawn about which procedures promise good performance. The steps

followed here are thus similar to those to be used in chapter 15: initial choice followed by later modification.

14.2 Threshold selection for the *CESAR4* data

The distribution of $\hat{f}(\tau)$ for the *CESAR4* dataset is shown in figure 14.2-1. The plot reveals generally right skewed distributions for all discriminant procedures[63].

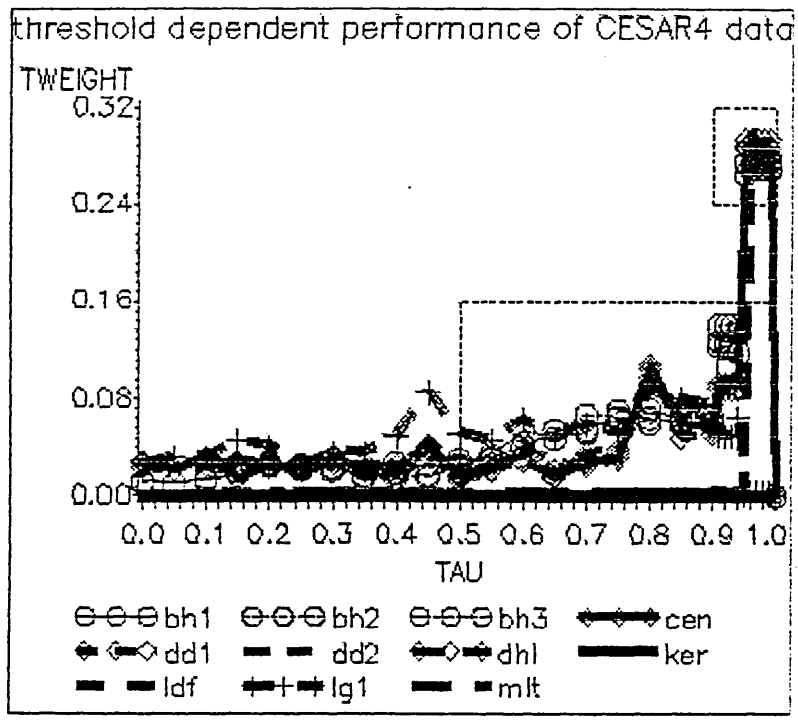


Figure 14.2-1: Distribution of $f(\tau)$

Looking at the upper right corner of the figure one can discern that the Bahadur procedures as well as the distance based procedures, respectively identifiable by the circle and diamond symbols, have the majority of their relative differences in posteriors at the upper end of the τ range. This in itself suggests good performance especially at higher thresholds. A strongly positively skewed distribution of $\hat{f}(\tau)$ implies that most allocations will be made with a high degree of confidence. By contrast the

63 The variable plotted vertically and labelled TWEIGHT stands for the estimated density $\hat{f}(\tau)$.

logistic procedure identifiable by the broken line and the plus symbols shows comparatively high frequencies of moderate to low relative posterior differences with a sudden dip down to zero for relative differences beyond $\tau = 0.95$. Figure 14.2-2 shows an exploded section of figure 14.2-1 to highlight differences at the top end of the τ range[64].

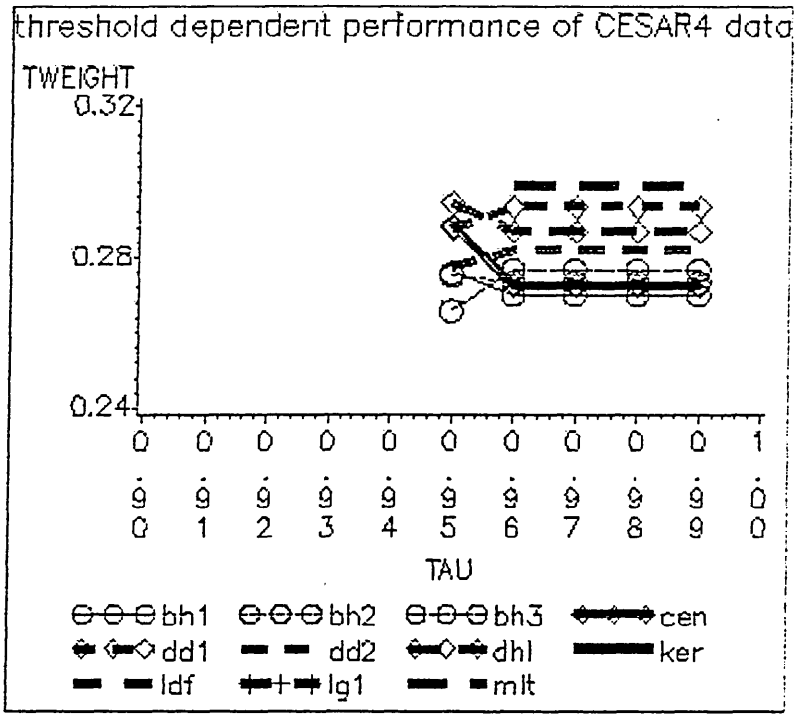


Figure 14.2-2: Blow up of figure 14.2-1

Figure 14.2-2 reveals that the centroid procedure has highest density for $\tau > 0.95$ followed by the *DHL* and *DD2* procedures, then the *BH2* and *DD1* and finally the *BH3* and *BH1* procedures. In terms of skewness, thus, choice of discriminant procedure would fall on these distance based procedures and on the family of Bahadur procedures for the *CESAR4* dataset.

[64] Note the changes in scaling of respective axes.

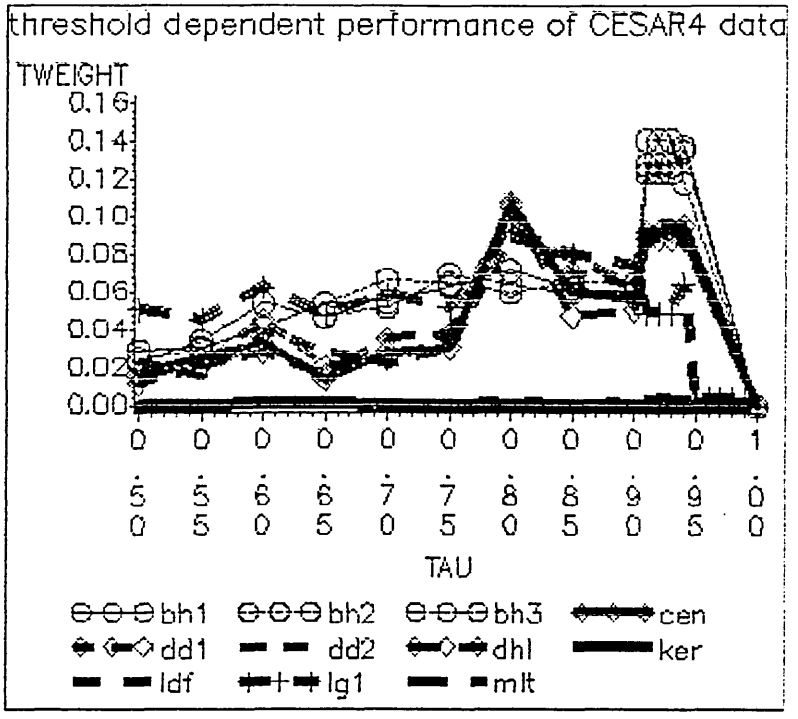


Figure 14.2-3: Blow up of figure 14.2-1

The next plot in figure 14.2-3 shows an exploded region of figure 14.2-1 for values of τ in the range $0.50 \leq \tau \leq 1.00$. It confirms again that the Bahadur procedures show considerable positive skew[65]. The curves for the distance based procedures are seen to lie below those for the Bahadur procedures but rise even more sharply for higher τ values. The expected prognosis of different procedures as inferred from the above analysis of the $f(\tau)$ distributions may be checked by inspecting figure 14.2-4 which displays the threshold dependent performance of $\varepsilon^{counting}$ under leaving-one-out crossvalidation conditions.

---

65 The fact that the lines for the Bahadur procedures "dip down" suddenly towards $f(\tau) = 0$ for $\tau = 1.00$ is not a contradiction to figure 14.2-1 but rather an artefact of the graphics software used: all points for which the ordinates lie within the Y-dimension of the graph are joined. Such instances are easily recognised by long straight lines in the graphs.

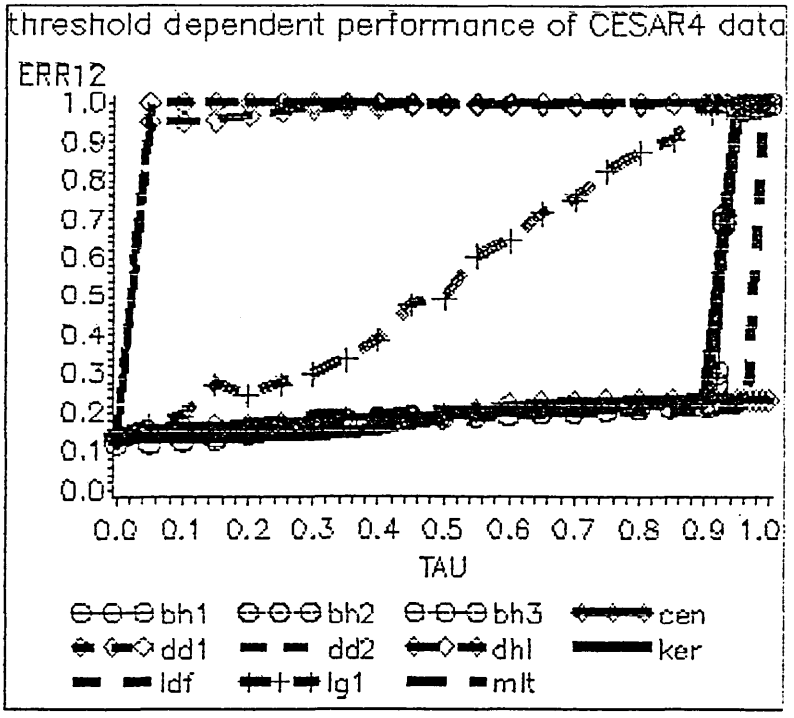


Figure 14.2-4: Leave-1-out based ε(count.) perf.

Ignoring the upper curves for the indirect procedures (*DD1* and *DHL*) because of poor performance, the marked continuous increase of $\varepsilon^{counting}$ for the logistic procedure immediately stands out. The lower curves on the other hand indicate comparatively good performance across a wide range of τ for the centroid, kernel, multinomial, linear discriminant and Bahadur procedures. Close inspection shows that the Bahadur procedures perform slightly better than the centroid procedure. Performance judged by $\varepsilon^{counting}$ thus to some extent[66] confirms possible candidates for initial choice on the basis of inspection of the $\hat{f}(\tau)$ distributions.

In chapter 9, section 4, several formalised approaches to procedure selection based on threshold dependent performance curves were suggested. One of these was the concept of *error doubling points*. Starting with $\tau = 0.00$ the point τ^{double} is determined for a given discriminant procedure as the point τ on the τ axis where the error rate doubles such that $\varepsilon(\tau^{double}) = 2\varepsilon(\tau=0.00)$. The procedure

[66] namely the Bahadur and centroid procedures

with the largest *error doubling point*, τ^{double}, performs best. Applying this logic to figure 14.2-4 shows however that it is difficult to discern between the slow rising curves for the Bahadur and centroid procedures. This formal approach also offers no obvious additional information beyond what can be directly extracted by visual inspection of the performance curves.

Detailed comprehensive graphs of performance plotted against classification threshold as in figure 14.2-4 have been included in appendix G. These were inspected in a similar fashion as adopted for figure 14.2-4 for $\varepsilon^{counting}$, $\varepsilon^{posterior}$ and η under leaving-one-out and hold-out crossvalidation conditions. The results of these graphical analyses are summarised in table 14.2-1.

Table 14.2-1 shows for each of three performance criteria the procedures selected on the basis of an analysis of the remaining performance curves plotted against relative posterior difference τ in appendix G. The curves are judged in terms of skew of the $\hat{f}(\tau)$ distributions, absolute level of the performance criteria φ, and by rate of change in the first half of the τ range. Two crossvalidation techniques are compared. Inspection of respective plots of threshold dependent performance for the *CESAR4* data in appendix G yields the conclusions summarised in table 14.2-1.

		$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
skew[67]	-	*BAH CEN DD1 DD2 DHL*		
leave-one-out	level	*KER LDF CEN MLT*	*LDF*	*LDF*
	change	*KER LDF CEN MLT*	*KER LDF CEN MLT*	*KER LDF CEN MLT*
hold-out	level	*LDF KER*	*LDF* *[68]	*LDF* *
	change	*KER LDF CEN MLT*	*KER LDF CEN MLT*	*KER LDF CEN MLT*

Table 14.2-1: Threshold analysis for *CESAR4* data

For the *CESAR4* data the kernel, linear discriminant and centroid procedure perform well. The indirect distance based procedures are poor while the logistic procedure exhibits intermediary performance. It is recommended to use threshold dependent performance when stability of estimates and reliability of a discriminant are important. The conclusions reached under leaving-one-out crossvalidation agree well with hold-out crossvalidation conditions.

Sections 14.3, 14.4 and 14.5 next describe procedure selection based on graphical threshold dependent performance analysis for the *NORMAL16*, *GRADE* and *IRIS* datasets following the same system as above.

14.3 Threshold selection for the *NORMAL16* data

Figure 14.3-1 shows the distributions of $\hat{f}(t)$ for all discriminant procedures[69]. The distributions are very similar in the case of the *NORMAL16* dataset when compared with the corresponding plot for the *CESAR4* data. To ease

---

[67] In terms of skew the listed procedures are expected to perform well for all performance criteria.

[68] The asterisk indicates that the linear discriminant is clearly superior to the rest.

[69] The variable plotted vertically and labelled TWEIGHT stands for the estimated density $\hat{f}(t)$.

interpretation again the two regions marked in figure 14.3-1 have been exploded in figures 14.3-2 and 14.3-3 respectively.

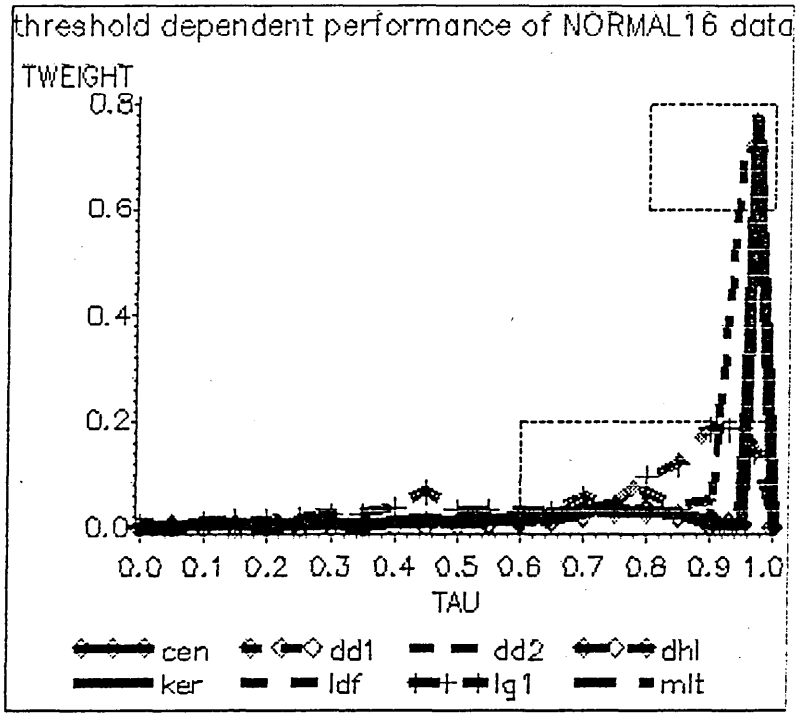


Figure 14.3-1: Distribution of $f(\tau)$

Figure 14.3-2 shows the upper region for $\tau \geq 0.80$. The plot reveals a high density of large posterior differences for the distance based procedures centroid, *DD1* and *DHL*. Figure 14.3-3 highlights differences for lower densities at $\tau \geq 0.60$. From this only the logistic stands out with substantial positive skew. Consequently the above procedures promise better performance judged by the $f(\tau)$ distributions.

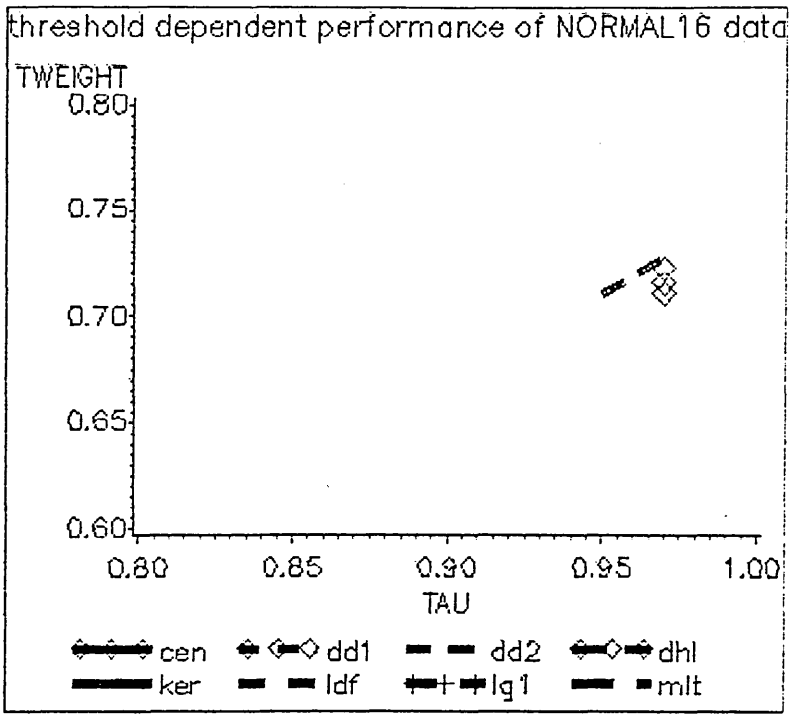


Figure 14.3-2: Blow up of figure 14.3-1

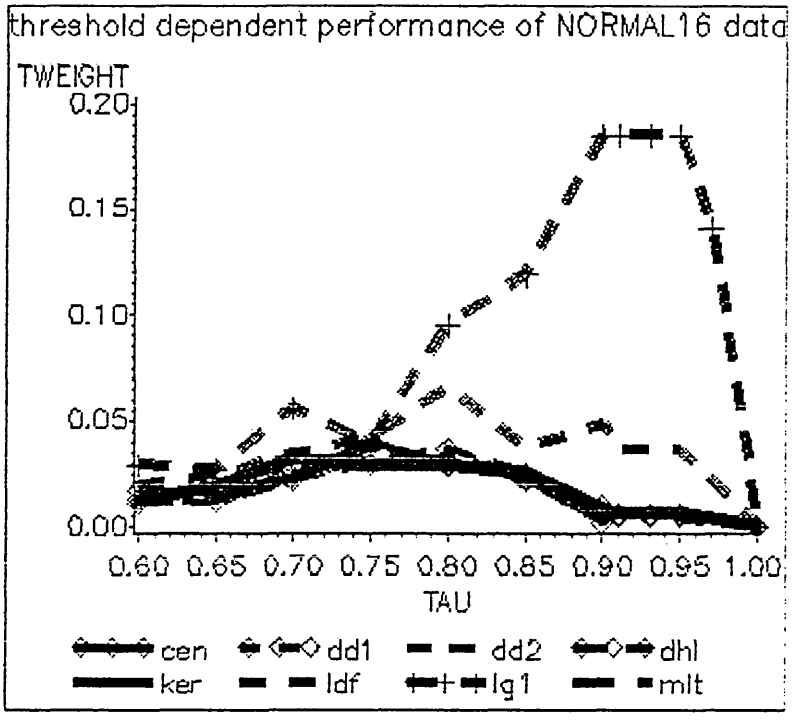


Figure 14.3-3: Blow up of figure 14.3-1

It may surprise that the *LDF* does not clearly outperform the other procedures considering that the *NORMAL16* dataset

is derived from a bivariate continuous normal distribution[70]. To understand this it is important to be aware of the "very" discrete nature of this dataset consisting of only $s = 10$ states. To illustrate this the data has been plotted in figure 14.3-4.

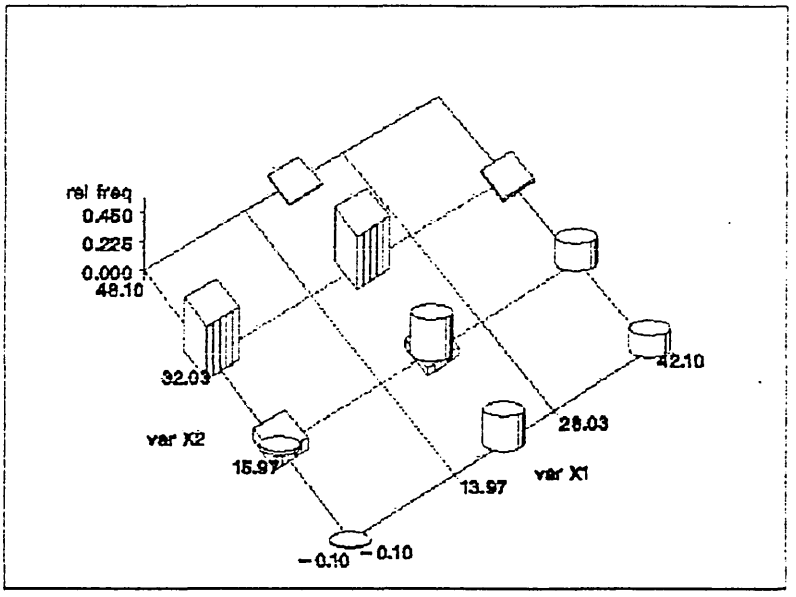


Figure 14.3-4: Plot of the *NORMAL16* data

The plot of the data shown in figure 14.3-4 highlights the strongly *discretised* bivariate nature of originally normally distributed data (chapter 11).

[70] see chapter 11

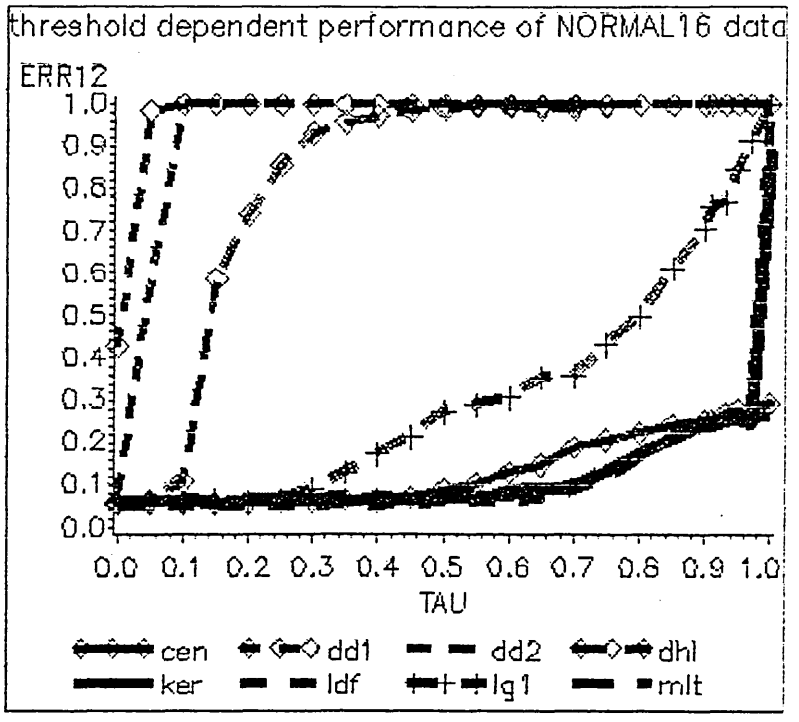


Figure 14.3-5: Leave-1-out based ε(count.) perf.

Inspection of the leave-1-out crossvalidated behaviour of $\varepsilon_{counting}$ (figure 14.3-5) confirms that the centroid discriminant performs well over the τ range. The continuous line with the diamond symbols for this procedure (*CEN*) is among the group of lines at the bottom of the plot all sharing similar low error rates of about *5* percent for values of $\tau \leq 0.50$ after which it rises slightly. The plot, however, also shows that the indirect *DD1*, *DD2* and *DHL* procedures perform worst. Their performance curves, conversely, rise sharply to reach almost maximum misallocation error rates for low values of $\tau \leq 0.20$. A close look at the lowest curve in this sense reveals a consistently good performance of the linear discriminant, which is what would have been expected from the construction of this artificial dataset. The other relevant plots of appendix G confirm this finding under leaving-one-out crossvalidation conditions. Under hold-out crossvalidation the linear discriminant shares its top position with the kernel, centroid and multinomial. These results are summarised in table 14.3-1.

		$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
skew[71]	-	*CEN DD1 DHL LG1*		
leave-one-out	level	*LDF* *[72]	*LDF*	*LDF*
	change	*KER CEN LDF MLT LG1*	*KER CEN LDF MLT LG1*	*KER CEN LDF MLT LG1*
hold-out	level	*KER CEN LDF MLT*	*KER CEN LDF MLT*	*KER CEN LDF MLT*
	change	*KER CEN LDF MLT LG1*	*KER CEN LDF MLT LG1*	*KER CEN LDF MLT LG1*

Table 14.3-1: Threshold analysis for *NORMAL16*

For the *NORMAL16* data the linear discriminant function, but also the centroid and multinomial procedures perform well. The logistic and the indirect distance based procedures are ruled out. The centroid perhaps does well because of the "centroid" nature of the artificial data derived from the normal distribution.

14.4 Threshold selection for the *GRADE* data

Figure 14.4-1 shows the distributions of $\hat{f}(\tau)$ for all discriminant procedures[73] applied to the *GRADE* dataset.

---

[71] In terms of skew the listed procedures are expected to perform well for all performance criteria.

[72] The asterisk indicates that the linear discriminant is clearly superior to the rest.

[73] The variable plotted vertically and labelled TWEIGHT stands for the estimated density $\hat{f}(\tau)$.

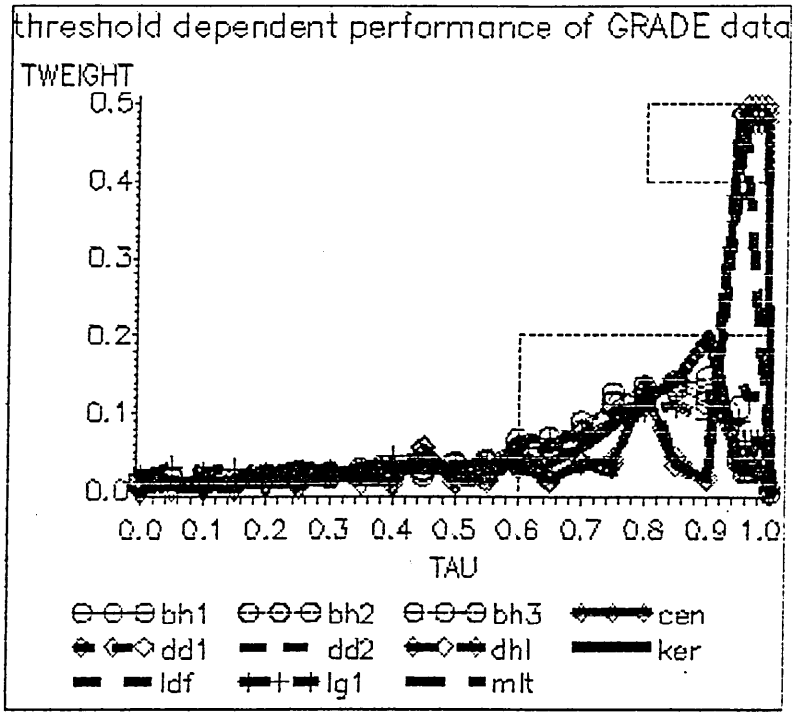


Figure 14.4-1: Distribution of $f(\tau)$

Again two regions have been singled out in figure 14.4-1 that are next magnified in figures 14.4-2 and 14.4-3 in order to reveal the detailed nature of skewness for different procedures. Figure 14.4-2 shows that the *DHL*, *DD1* and *CEN* procedures exhibit greatest degree of positive skew.

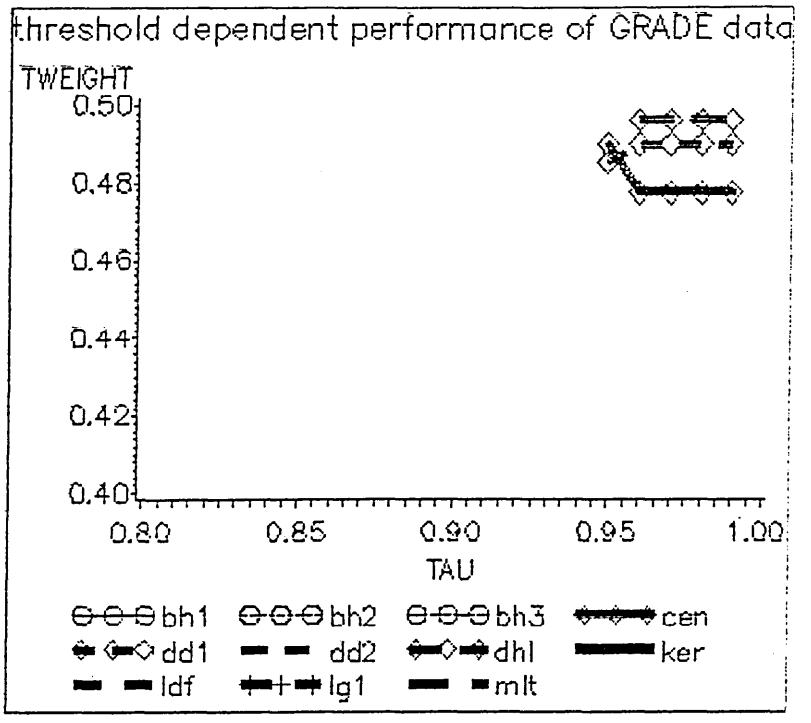


Figure 14.4-2: Blow up of figure 14.4-1

Figure 14.4-3 shows that the kernel and linear discriminant stand out slightly among the procedures with less positive skew.

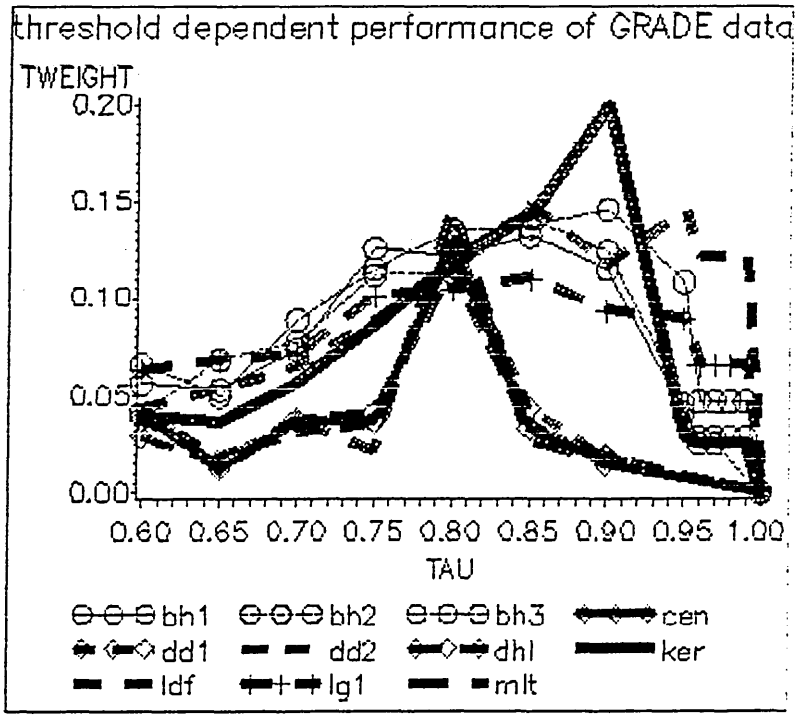


Figure 14.4-3: Blow up of figure 14.4-1

The visual inspection of performance curves for the *GRADE* data in figure 14.4-4 clearly enables identification of two groups of discriminant procedure. Of these the range of Bahadur procedures, the multinomial, kernel and linear discriminant show better performance. For threshold values up to about $\tau = 0.70$ the Bahadur procedures are best in this group.

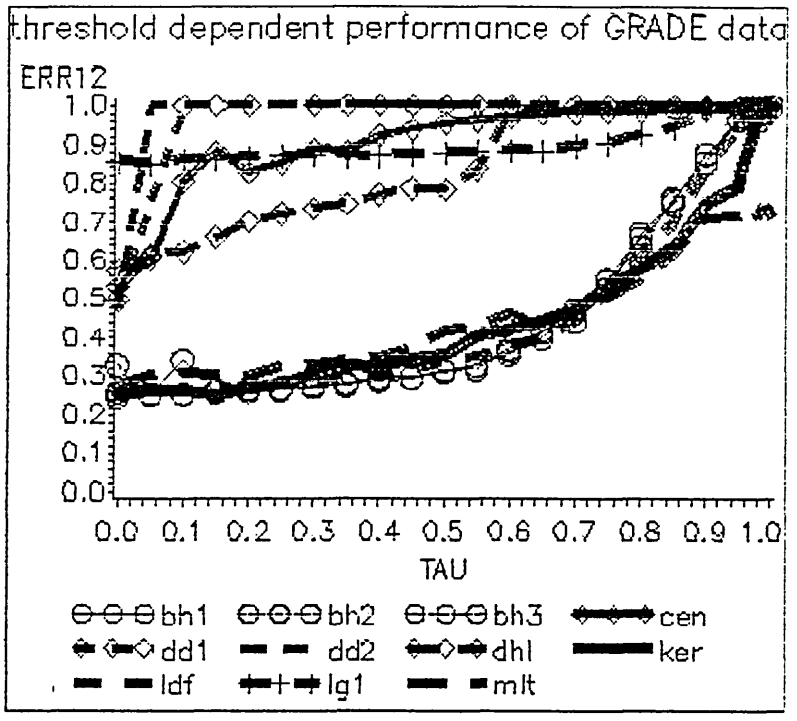


Figure 14.4-4: Leave-1-out based ε(count.) perf.

Table 14.4-1 shows for each of the three performance criteria the procedures selected on the basis of an analysis of the remaining performance curves plotted against relative posterior difference τ in appendix G. The curves are judged in terms of skewness of the $\hat{f}(\tau)$ distributions, absolute level of the performance criteria φ, and by rate of change in the first half of the τ range. Two crossvalidation techniques are compared. Inspection of respective plots of threshold dependent performance for the *GRADE* data in appendix G yields the conclusions summarised in table 14.4-1.

		$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
skew[74]	-	*DHL DD1 CEN*, & then *KER LDF*		
leave-one-out	level	*BAH MLT KER LDF*	*BAH MLT KER LDF*	*BAH MLT KER LDF*
	change	*BAH MLT KER LDF*	*BAH MLT KER LDF*	*BAH MLT KER LDF*
hold-out	level	*BAH MLT KER LDF*	*BAH MLT KER LDF*	*BAH MLT KER LDF*
	change	*BAH LDF*	*BAH LDF*	*BAH LDF*

Table 14.4-1: Threshold analysis for *GRADE* data

14.5 Threshold selection for the *IRIS* data

Figure 14.5-1 shows a right skewed distribution of $\hat{f}(\tau)$ for all discriminant procedures[75] for the *IRIS* dataset.

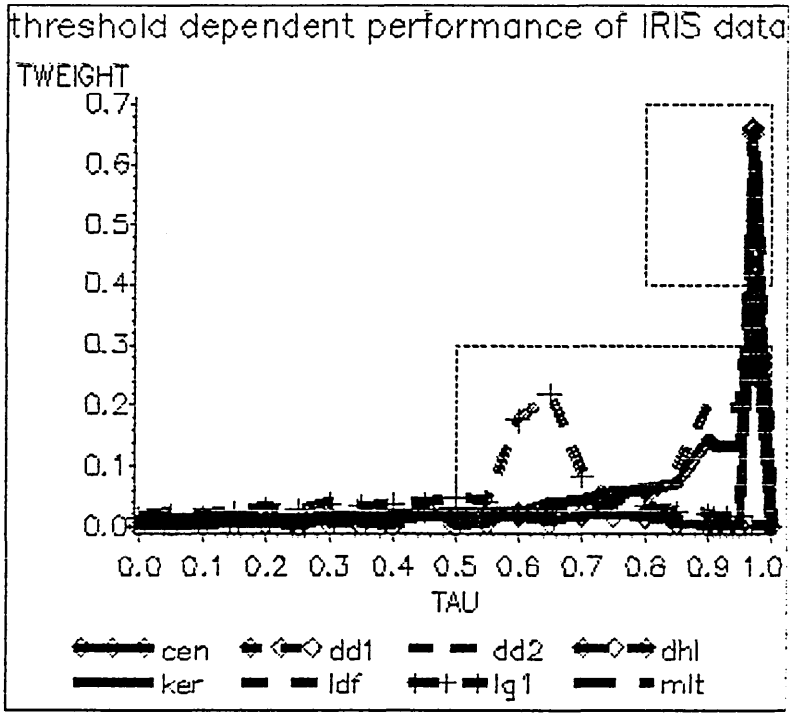


Figure 14.5-1: Distribution of $f(\tau)$

---

[74] In terms of skew the listed procedures are expected to perform well for all performance criteria.

[75] The variable plotted vertically and labelled TWEIGHT stands for the estimated density $\hat{f}(\tau)$.

Corresponding enlarged regions of figure 14.5-1 are displayed in figures 14.5-2 and 14.5-3 respectively.

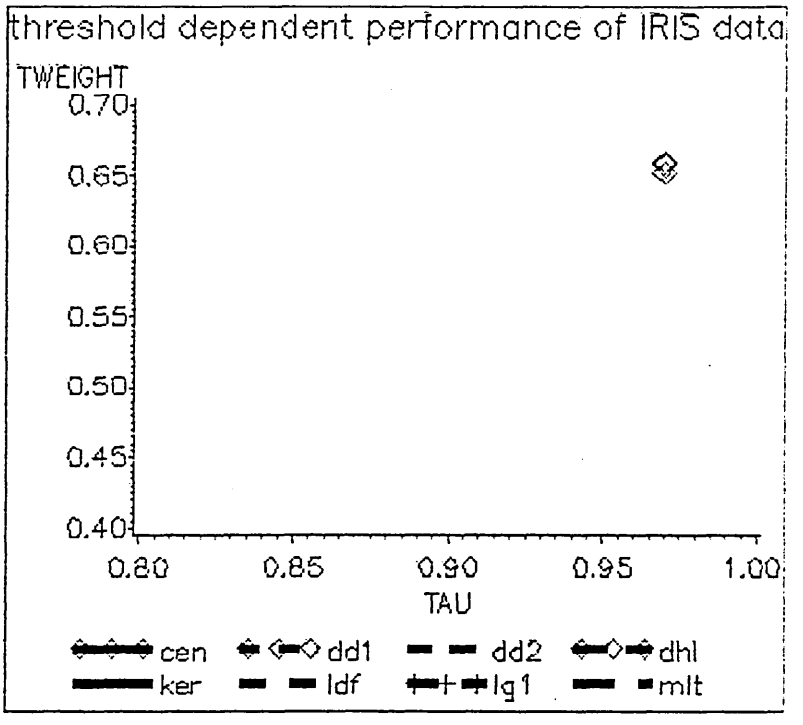


Figure 14.5-2: Blow up of figure 14.5-1

Figure 14.5-2 highlights the strongest positive skew for the *CEN*, *DD1* and *DHL* procedures.

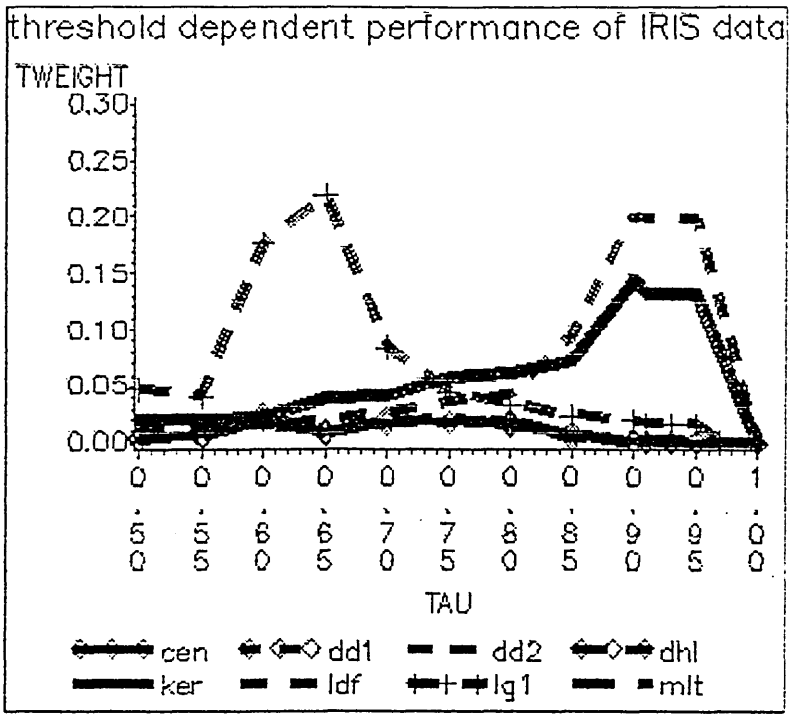


Figure 14.5-3: Blow up of figure 14.5-1

Figure 14.5-3 shows that the linear discriminant procedure and the kernel procedure also have considerable densities of relative posterior differences at the upper end of the τ range.

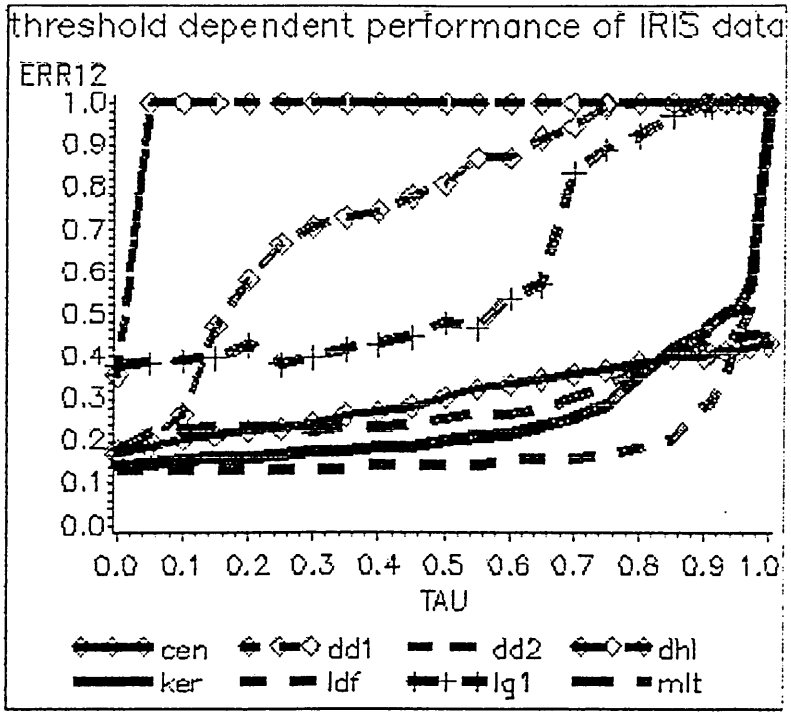


Figure 14.5-4: Leave-1-out based ε(count.) perf.

The analysis of threshold dependent performance for $\varepsilon_{counting}$ in figure 14.5-4 reveals a consistently superior performance for the linear discriminant across *95* percent of the τ range.

Table 14.5-1 shows for each of three performance criteria the procedures selected on the basis of an analysis of the performance curves plotted against relative posterior difference τ in appendix G. The curves are judged in terms of the skewness of the $\hat{f}(\tau)$ distributions, absolute level of performance criteria φ, and by rate of change in the first half of the τ range. Two crossvalidation techniques are compared. Inspection of respective plots of threshold dependent performance for the *IRIS* data in appendix G yields the conclusions summarised in table 14.5-1.

		$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
skew[76]	-	*CEN DD1 DHL*, & then *LDF KER*		
leave-one-out	level	*LDF*	*LDF*	*LDF*
	change	*LDF KER MLT*	*LDF KER MLT*	*LDF KER MLT*
hold-out	level	*LDF KER*	*LDF*	*LDF*
	change	*LDF KER*	*LDF*	*LDF*

Table 14.5-1: Threshold analysis for *IRIS* data

Table 14.5-1 reveals with respect to threshold dependent behaviour of performance criteria that the linear discriminant generally is the best discriminant for this dataset. Considering the original shape of this dataset this might have been expected. The analysis of skew of the $f(\tau)$ distributions by contrast does not initially lead to the same choice.

14.6 Conclusions

Assessment of performance of discriminant procedures using analysis of classification thresholds is based on inspection of the distributions[77] of relative differences in posteriors, $f(\tau)$, and on inspection of the values of performance criteria at different classification thresholds. With respect to the latter the absolute level of values of performance criteria as well as the rate of change are the major characteristics on which judgement should be made. Using this approach selections of optimal discriminant procedures is possible for the three real datasets (*CESAR4*, *GRADE* and *IRIS*) and the one artificial (*NORMAL16*) dataset considered in the examples. There is

[76] In terms of skew the listed procedures are expected to perform well for all performance criteria.

[77] see chapters 9 and 10 for information on the way in which the $f(\tau)$ distributions are estimated

also generally little difference in selections between the leaving-one-out and the hold-out crossvalidation technique.

In terms of the $f(\tau)$ distribution for the *CESAR4* data choice falls on the Bahadur, centroid and on the nonparametric distance based procedures. In terms of actual estimates of performance criteria the kernel, the linear discriminant and the centroid procedure are superior. With respect to $\varepsilon^{posterior}$ and η the linear discriminant function performs particularly well.

In terms of the $f(\tau)$ distribution for the *GRADE* data choice falls on the nonparametric distance based procedures and, to a lesser extent, on the kernel and linear discriminant procedures. In terms of actual values of performance criteria the Bahadur, multinomial, kernel and linear discriminant procedures are superior.

As illustrated in chapter 9 variable thresholds are equivalent to fixed thresholds when discrimination is between $g = 2$ populations. When there are $g \geq 3$ populations it was shown that there are advantages in using variable classification thresholds as opposed to fixed thresholds. Although not demonstrated here explicitly this will in the case of the *GRADE* data have lead to more stable allocations because here discrimination is between $g = 3$ populations.

In terms of the $f(\tau)$ distribution for the *IRIS* data choice falls mainly on the linear discriminant and next on the kernel and nonparametric distance based procedures. In terms of thresholded performance the linear discriminant is clearly singled out by all performance criteria.

In terms of the $f(\tau)$ distribution for the *NORMAL16* data choice falls on all but the first order logistic and the multinomial procedures. In terms of actual values of performance criteria the linear discriminant but also the kernel, centroid and multinomial procedures perform comparatively well. The superior performance of the linear

discriminant function is particularly marked under leaving-one-out crossvalidation conditions.

Generally sufficient information may be gleaned from visual inspection of the estimated densities of $f(\tau)$ and the plots of threshold dependent performance for a given crossvalidation technique. It must be emphasised that graphical analysis of these threshold dependent performance curves very rapidly allows the rejection of unacceptable discriminant procedures. In all of the examples the differentiation of procedures in terms of non-thresholded performance is considerably more difficult than even at only slightly raised thresholds.

For this reason it is recommended that it is sufficient to concentrate on the lower end of the τ-range below values, of approximately *0.30* when using classification thresholds. This is important, because it can reduce the computing effort by for instance choosing a rougher metric for the upper end of the τ-range.

If possible, leaving-one-out crossvalidation should be used to allow better resolution of the curves when hold-out techniques lead to similar curves.

Beyond visual inspection there appears to be no further need for formalising the analysis of performance curves by introduction of concepts such as *error doubling points* as discussed in chapter 9.

I: INTRODUCTION

II: REVIEW

III: METHOD

IV: RESULTS

13. Analysis of Performance Criteria	14. Analysis of Classification Thresholds

15. Application of Selection Rules

15.1 Strategy of procedure selection

15.2 Procedure choice for the *CESAR4* data

15.3 Procedure choice for the *CREDIT* data

15.4 Procedure choice for the *CHD* data

15.5 Procedure choice for the *BANANA* data

15.6 Procedure choice for the *MA4353* data sets

15.7 Procedure choice for the *INTERAC1* data

15.8 Procedure choice for the *IRIS* data

15.9 Procedure choice for the *EDUC* data

15.10 Conclusions

V: DISCUSSION

The aims of this chapter are twofold: (1) demonstration of how to apply the selection tree developed in chapter 12, and (2) demonstration of the validity of using a selection tree. The general strategy for application of selection trees is outlined in section 15.1 while actual applications to the datasets are given in sections 15.2 to 15.9.

15.1 Strategy of procedure selection

In order to show the usefulness of a selection tree it is applied to a selection of real and artificial datasets. The question of whether initial procedure choice is matched by correspondingly good performance as measured by various criteria of performance will be of particular interest. The performance estimates presented in chapter 13 will be used to test the validity of initial procedure choice. Results are summarised in tables giving the procedures for which favourable values of performance criteria are computed. These summaries are condensed from individual results tabulated by dataset and discriminant procedure in appendix A, B and C.

To allow assessment of generalisability[78] of the selection tree to future applications, datasets exhibiting a variety of different data structures have been chosen. The paradigm for procedure selection as developed in section 12.1 of chapter 12 is the starting point of the procedure selection process. With reference to the figure 12.1-2 the *skill/experience* and *constraints* factors are considered of secondary importance while the emphasis is initially on the *data* and *model*, and later on the *demand* factors. The selection tree of figure 12.6-3 in chapter 12 is used in the following demonstrations for real and artificial

[78] but see also caveat at end of chapter 11

datasets in sections 15.2 to 15.9. The discussion for the *CESAR4* dataset in section 15.2 and for the *IRIS* dataset in 15.8 also includes a comparison between thresholded and non-thresholded performance analysis. For reference the selection tree from chapter 12 is repeated below in figure 15.1-1. The three primary factors influencing choice *model*, *data* and *demands* have been positioned (within brackets) in figure 15.1-1.

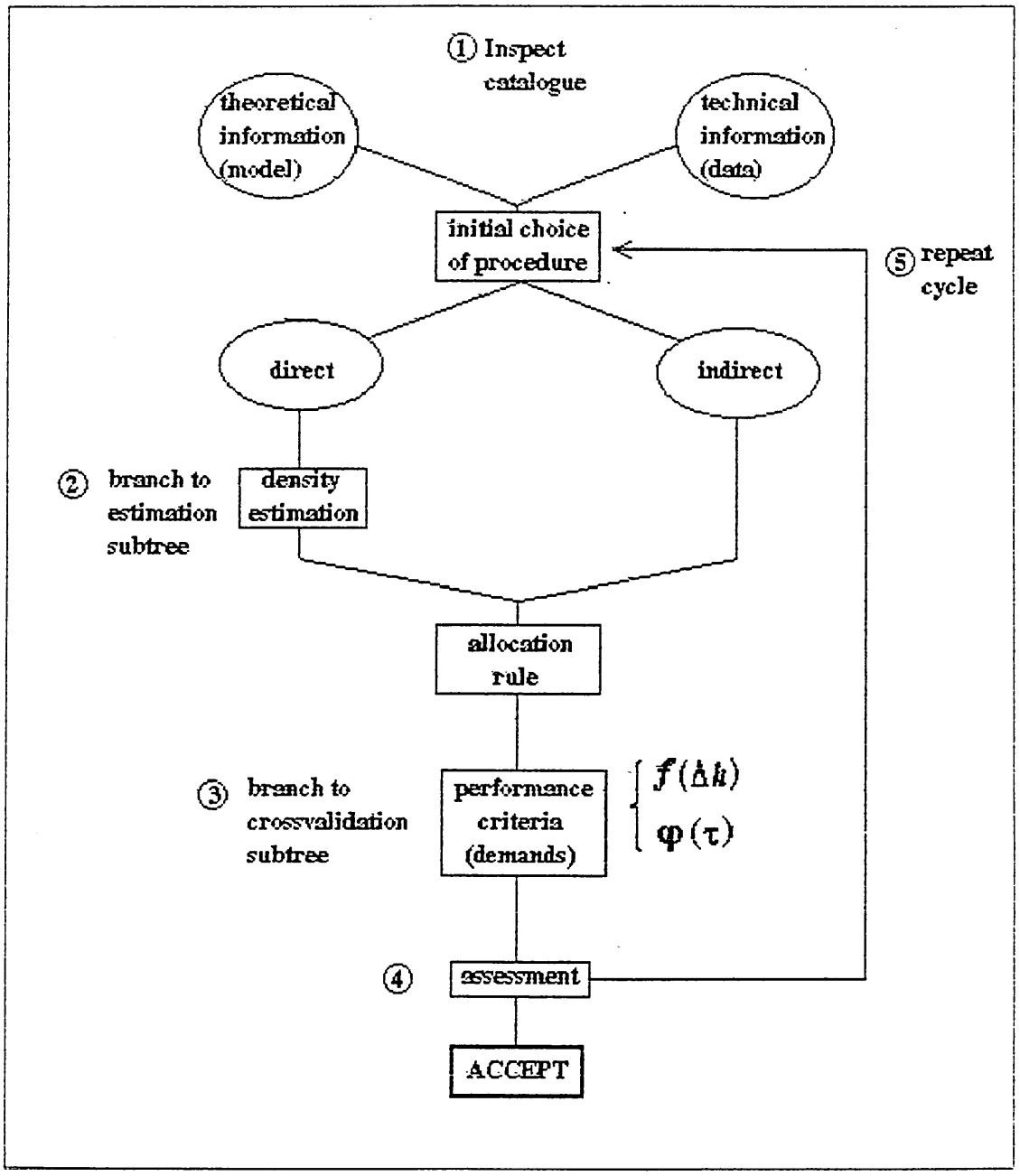


Figure 15.1-1: Procedure selection tree

Performance is judged on three key features: (1) hold-out based expectations, (2) expected standard errors and (3) bias based on unconditional estimates[79]. When considering the *model* and *data* factors frequent reference will be made to the corresponding tables of characteristics of respective datasets in chapter 11. These tables include descriptive statistics of group specific means, standard deviations and correlation matrices as well as in some cases condensed results from loglinear analyses carried out on the datasets prior to running the discriminant analyses.

15.2 Procedure choice for the *CESAR4* data

The *CESAR4* data is a medium sized sample of real data with dichotomous predictors and a dichotomous response.

Technical and theoretical admissibility: Beginning with *technical* admissibility (the right hand *data* branch of number 1 in figure 15.1-1) the features of this dataset (chapter 11) do not preclude any of the eleven discriminant procedures (*BH1, BH2, BH3, CEN, DD1, DD2, DHL, KER, MLT, LDF, LG1*) hence general admissibility is given. A restriction would apply if some of the predictors X_j were not dichotomous such that the Bahadur procedures could not be computed (chapter 4). On *theoretical* grounds (left hand *model* branch of number 1) the multivariate dichotomous nature of the data would imply any of the Bahadur models as these are specifically designed for such distributions. The multinomial model must also be considered because it makes no assumptions about the individual variables and their interactions. Due to non-normality of the multivariate dichotomous distributions the linear discriminant function might be ruled out on theoretical grounds. The robustness of the linear discriminant however is an argument in favour

[79] It was seen in chapter 13 that estimates of bias of the performance criteria depended on whether baseline values of performance criteria were computed using conditional or unconditional estimates. For this reason the conditional estimates are not used.

of the *LDF*. It was seen in section 5 of chapter 13 that performance of the *LDF* broke down noticeably only when the number of discrete states dropped to very low levels. The *CESAR4* data has $s = 12$ states which might therefore be enough to support reasonable performance of the *LDF* model.

The question as to what medical factors are likely to influence the rate of caesarean sections arose because of a growing concern about the continuously increasing trend that had been observed by perinatal surveys since the middle 1970's. The prior probability of $\pi = 0.146$[80] reflects the caesarean section rate in a sample of the German[81] population in 1986. The *CESAR4* data shows the caesarean section rate as a function of the dichotomous variables "position of fetus in womb", "twin pregnancy", "previous caesarean section" and "placental insufficiency". In absence of knowledge about a statistical model that underlies the *CESAR4* data, inspection of the variables and their likely interrelation on medical grounds is carried out (still number 1 in figure 15.1-1). From the summary table 11.1-3 in chapter 11 it may be seen that the relative differences in the means for the four predictor variables lie between about *80* and *100* per cent. The standard deviations on the other hand are closer, especially when considering the different sample sizes. Still these features suggest that a reasonable separation of the two populations should be possible. On medical grounds one might expect a correlation between position of fetus in the womb and twin pregnancy and also a smaller one between twin pregnancies and placental insufficiency and finally perhaps an even smaller one between previous caesarean or uterus surgery and placental insufficiency. The correlation matrices R_1 and R_2 in chapter 13, however, reveal only one small, yet statistically significant, correlation of *0.147* and *0.115* respectively between twin pregnancy and previous caesarean section or uterus surgery. This contradicts theoretical expectations. The expected correlation between

---

[80] see summary table 11.1-3 in chapter 11.

[81] in the state of Lower Saxony (Niedersachsen).

position and twin pregnancy is significant only in the pooled sample. The loglinear analysis yields *4* significant main effects quite in keeping with the relatively large univariate differences in the means (see above).

Initial choice of procedure: The above implies a higher order model, such as a *BH2* or even the *BH3*. The *LDF* might also be expected to perform reasonably well.

Density estimation: This follows the parametric techniques as outlined in chapter 4.

Crossvalidation of performance criteria: The sample size $n = 1544$ is comparatively large and so the more intensive crossvalidation techniques such as leaving-v-out or even leaving-one-out are not used in favour of hold-out crossvalidation.

Assessment: Table 15.2-1 has two parts. The upper section above the double horizontal lines shows procedure choices on the basis of using hold-out based non-thresholded estimates of performance (from appendix A, C and E.) for $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η in terms of expectation $E[\varphi]$, standard error $se(\varphi)$ and unconditional bias $ub(\varphi)$[82]. The lower section is a repetition of the corresponding summary table from chapter 14 giving details of procedure choice based on threshold dependent performance curves for assessments of procedure reliability.

For the *CESAR4* data, table A-1 from appendix A gives non-thresholded hold-out estimates for $\varepsilon_{counting}$ of *0.123* for both the *BH1* and the *BH3* discriminant procedures. All tables in the appendix have been rounded to *3* significant post decimal digits. In the above case the actual values of $\varepsilon_{counting}$ for the *BH1* and *BH3* procedures are in fact *0.1234* and *0.1229* respectively. Hence only the *BH3* procedure is entered in the upper left cell of table 15.2-1. In all

[82] The symbol φ is short for $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η.

subsequent tables summarising procedure choice the actual data[83] have therefore been consulted in cases of doubt. The corresponding best - and unambiguous - estimate from table A-2 for $\varepsilon_{posterior}$ is *0.072* for the *LDF*. As this performance is distinctly better than for the other procedures this entry in the upper middle cell of table 15.2-1 is additionally marked with an asterisk. The best η value (*0.897*) is again exhibited for the *LDF* procedure yielding the entry in the upper right cell. The second line of table 15.2-1 shows the procedure for which the estimate of the unconditional standard error is at a minimum for the three performance criteria. The actual values are *0.0015*, *0.0036* and *0.0021*, respectively. In each case the *BH3* procedure has the smallest value. The third line shows procedures for which estimates of the absolute value of the unconditional bias are a minimum. Actual values accurate to *4* significant post decimal digits are *0.5728*, *0.0000* and *0.0569*, respectively.

statistic		$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
expectation uncond s.e. uncond bias		*BH3* *BH3* *BH3*	*LDF**[85] *BH3* *BH1*	*LDF** *BH3* *BH1*
skew[84]	-	*BAH CEN DD1 DD2 DHL*		
leave-one-out	level	*KER LDF CEN MLT*	*LDF*	*LDF*
	change	*KER LDF CEN MLT*	*KER LDF CEN MLT*	*KER LDF CEN MLT*
hold-out	level	*LDF KER*	*LDF**[86]	*LDF**
	change	*KER LDF CEN MLT*	*KER LDF CEN MLT*	*KER LDF CEN MLT*

Table 15.2-1: Procedure choice for *CESAR4* data

---

[83] available from the author

[84] In terms of skew the listed procedures are expected to perform well for all performance criteria.

[85] The asterisk indicates clearly superior performance for the given procedure.

[86] The asterisk indicates that the linear discriminant is clearly superior to the rest.

Inspection of table 15.2-1 shows that choice of procedure will initially depend on whether non-thresholded performance ($\tau = 0.00$) or thresholded performance ($0.00 < \tau \leq 1.00$) is used as a yardstick. This difference is particularly noticeable for the misallocation error, $\varepsilon_{counting}$. While the upper part of table 15.2-1 points to the Bahadur procedures on the basis of non-thresholded performance, the lower part lists the kernel, linear discriminant, centroid and multinomial procedures. To a lesser extent differences are also seen for $\varepsilon_{posterior}$ and η. The reasons for these differences in choice result from the fact that in one case performance is judged solely for $\tau = 0.00$ while in the other the response of performance over a range of values for τ is considered.

Recommendations for final choice: The above highlights the consequences that the choice of performance criteria in turn can have on subsequent choice of discriminant procedure. It also illustrates the point made in chapter 12 that the demands expected of a discriminant procedure have to be specified in advance. Table 15.2-1 also confirms the initial choice of the linear discriminant procedure as a potential candidate for the *CESAR4* data. This is the case when performance is judged in terms of non-thresholded values of $\varepsilon_{posterior}$ or η. Note that here this also holds for thresholded performance for these posterior probability based performance criteria in terms of level.

The above result is a typical illustration of the case where a theoretically inadmissible procedure (the linear discriminant) may lead to better performance than a procedure based on the correct model. In the terminology developed in chapter 7 this is equivalent to stating that performance judged by $\varphi(\hat{\delta},\hat{f},(\hat{\theta}))$ is superior to performance judged by $\varphi(\hat{\delta},f,(\hat{\theta}))$. The difference between the first expression and the last one is that the inadmissible (wrong, because linear discriminant) model assumption in the former is indicated by the estimated underlying

distribution $\hat{f}$ while in the latter the presumed admissible (i.e. the third order Bahadur) is indicated by f^{87}.

If good non-thresholded performance in terms of low $\varepsilon^{counting}$ on the basis of the present sample is considered more important than reliability of the discriminant in terms of low $\varepsilon^{posterior}$ or high η then choice would fall on the third order Bahadur procedure. If priorities are converse then choice would fall on the linear discriminant procedure.

For practical purposes two situations may be distinguished (a) action on the basis of the given data and (b) collection of further data with the intention of narrowing down choice of data model. Considering the former clearly (see chapter 4) the linear discriminant is an easier procedure to use under realistic conditions (e.g. antenatal clinic) because the linear predictor may be computed on most pocket calculators. By contrast the 3rd order Bahadur involves considerably more parameters and is more complex to handle for non-specialist staff. In addition the observed correlations, though significant, are comparatively low, thus casting doubt on the stability of such a high order interaction procedure. This potential susceptibility to sampling variation is also borne out by the threshold dependent performance analysis that entirely excludes all Bahadur procedures as good candidates, irrespective of crossvalidation technique and performance criterion. As reliability of a discriminant procedure routinely implemented for prediction of the risk of caesarean section delivery is presumably more important than low error rates, $\varepsilon^{posterior}$ and η are the crucial performance criteria. On the basis of these the linear discriminant is again suggested as optimal, especially in terms of level. With reference to situation (b) above it may be noted that the information contained in the

---

[87] With respect to figure 7.6-2 in chapter 7 one may conclude that the current sample size must lie to the left of the *reversal point, n^**.

variables X_1 to X_4 is routinely recorded in antenatal clinics. Thus increasing the sample size for future finer calibration of the underlying data model offers a further affordable means for improving procedure choice.

15.3 Procedure choice for the *CREDIT* data

The *CREDIT* data were chosen for illustration of the selection process because again it is a medium sized dichotomous response sample similar to the *CESAR4* dataset of the previous section, yet its particular feature is the set of interrelated demographic nominal and ordinal predictors.

Technical and theoretical admissibility: The ordinal nature of *4* (X_1, X_2, X_5, X_6) of the *6* the predictors precludes the range of Bahadur models in terms of technical admissibility. All other procedures may be applied. On data theoretic grounds all except the linear and quadratic discriminant functions are admissible. The ordinal scaling of the majority of predictors might suggest the direct entry of independent effects in the logistic model[88]. This variant of the logistic model is provided by some statistical packages and allows the modelling of ordinal predictors.

The *CREDIT* data has *236* different states which means that many of the cells have less than *10* or even zero observations. This rules out the multinomial model due to parameter abundance. Other nonparametric procedures such as the *DHL* and *DD2* might be expected to do fairly well. Due to the ordinal nature of most other predictors the linear discriminant function or the *QDF* should also do well especially as a consequence of robustness of the *LDF*. The same holds for the kernel procedure *KER* and possibly not so much for the centroid because of interaction effects

[88] see chapter 4.

leading to patterns in the data. The predictors of creditworthiness of bank customers (X_1 currently held account, X_2 past paying morale, X_3 savings, X_4 purpose of credit, X_5 assets and X_6 employment) lead one to expect considerable interaction between them. Duration of employment will be related to capital and this to past payment morale and presence of other savings etc. So higher order logistic models would be expected to perform well too.

Univariate analysis of the relative differences in means - analogous to the analysis carried out for the *CESAR4* example - reveals that the differences range from about *10* to *30* per cent for all variables except for X_3 (*presence of savings*). Here the univariate relative difference is *130* per cent. The loglinear analysis confirms the relevance of X_2 but also lists X_1, X_3 and X_4 as further significant main effects. Major significant, yet low correlations are seen between *past paying morale* and *assets* (*0.199* and *0.193* in Π_1 and Π_2 respectively), *current balance* and *employment* (*0.156* and *0.137*), *current balance* and *savings* (*0.118* and *0.123*) and *past paying morale* and *purpose of credit* (*-0.112* and *-0.100*). These correlations single out *current balance* and *past paying morale* as the chief potential predictors.

Initial choice of procedure: The distance based *DHL* or *DD2* procedures.

Density estimation: Not required for indirect procedures.

Crossvalidation of performance criteria: Medium to large sample size implies that hold-out crossvalidated estimates should give satisfactory results.

Assessment: The results for non-thresholded performance from corresponding tables in the appendix are summarised in table 15.3-1. The best performance estimates from appendix A are *0.200*, *0.167* and *0.772* respectively for $\varepsilon^{counting}$, $\varepsilon^{posteriro}$ and η. The first line of the table confirms

initial choice with respect to non-thresholded expectation of all performance criteria. In each case a distributional distance procedure performs best. The fact that on the other hand the linear discriminant and the logistic procedures perform better in terms of precision and bias suggests that the distance based procedures might be overfitted.

statistic	$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
expectation	*DD2*	*DD1*	*DD2*
uncond s.e.	*LDF*	*LDF*	*LDF*
uncond bias	*LG1*	*LDF*	*LDF*

Table 15.3-1: Procedure choice for *CREDIT* data

Recommendations for final choice: In the banking business errors of misallocation naturally have direct monetary consequences. Falsely assuming creditworthiness is disadvantageous because it will incur extra costs in retrieving capital laid out. On the other hand falsely assuming a customer to be not worthy of credit similarly means a loss of potential interest. The losses caused by incorrect decisions need not be the same in both these cases. In a sense therefore this is a typical example for considering differential misallocation costs[89].

Decisions in banking must be taken quickly, otherwise customers may go elsewhere. Therefore, although it may be comparatively inexpensive to obtain further data the updated discriminant procedure may arrive too late. Hence, resampling is ruled out. Thus the choice narrows down to a choice between the *DD1* or *DD2* and the *LDF* procedure on the basis of non-thresholded performance. In particular a bank manager prepared to take a certain risk for the sake of a higher yield would be recommended to choose the modified distributional distance procedure *DD2* while a more conservative manager should use the linear discriminant.

[89] but see also chapter 7

15.4 Procedure choice for the *CHD* data

The *CHD* dataset was selected because it features a comparatively low prior with $\pi_1 = 0.069$ for presence of coronary heart disease. The other characteristic feature is the ordinal nature of the two predictors: blood pressure measured at *4* levels and serum cholesterol content also measured at *4* levels.

Technical and theoretical admissibility: On theoretical grounds the linear and quadratic discriminant functions are not admissible in the strict sense for all discrete datasets. Yet as here two originally continuous and probably also normally distributed variables (systolic blood pressure and cholesterol level) are represented at *4* ordinal levels one may nevertheless expect a slightly better performance for the linear discriminant. On medical grounds one would expect an interaction effect, although the correlation should be higher for diastolic blood pressure. The corresponding correlation matrix from chapter 11, however, does not show up these effects ($\rho_{12} = 0.095$). The absence of a substantial interaction[90] suggests using a parsimonious procedure such as the multinomial or the independent *BH1* model. When considering the low correlation between the predictor variables jointly with the relative differences in means (only about *20* per cent) it becomes clear that separation of the populations will not be good. As ordinal predictors begin to approach continuous distributions one can expect other procedures apart from the linear discriminant function to improve. This implies the kernel density based discriminant procedure and also the indirect centroid procedure.

Initial choice of procedure: The multinomial or *BH1* discriminant procedure.

Density estimation: As specified in chapter 4.

---

[90] although significant at the 5 per cent level a Kendall's tau_b of 0.095 must be considered small.

Crossvalidation of performance criteria: The sample size is comparatively small considering the number of discrete states. This therefore suggests using a more elaborate form of crossvalidation such as leaving-one-out to reduce bias.

Assessment: Non-thresholded performance is given in table 15.4-1. Corresponding estimates of expected values for the *MLT*, *DHL* and *KER* procedures from appendix A are *0.069*, *0.068* and *0.931* respectively. Although these values are good in absolute terms it turned out that the misallocation rate is of the same magnitude as the prior probability for π_1. In fact π_1 was entirely classified as π_2 by the discriminant procedure with the smallest error rate, $\varepsilon^{counting}$. The *CHD* data are a good example for using differential misallocation costs, especially as sensitivity for the medical condition must be given a high priority.

statistic	$\varepsilon^{counting}$	$\varepsilon^{posterior}$	η
expectation	*MLT*	*DHL*	*KER*
uncond s.e.	all procs*[91]	*LDF*	*LDF/KER*
uncond bias	*KER/LDF*	*MLT*	*KER*

Table 15.4-1: Procedure choice for *CHD* data

Recommendations for final choice: The ambiguous results suggest adjusting for different misallocation costs and perhaps adopting a selective discrimination approach (see chapter 12).

15.5 Procedure choice for the *BANANA* data

The artificial *BANANA* dataset was constructed to demonstrate that a discriminant procedure based on a straight separation line would not perform well. The

[91] The estimates for the unconditional standard errors are virtually equal for all admissible procedures. Thus no single one stands out among the rest with respect to precision.

distinct structure of the data is also expected to exclude other procedures. The data consists of separately sampled artificial bivariate data with priors $\pi_i = 0.500$ and two uncorrelated ordinal predictors with *6* levels each. The data is plotted in *3* dimensions in figure 15.5-1.

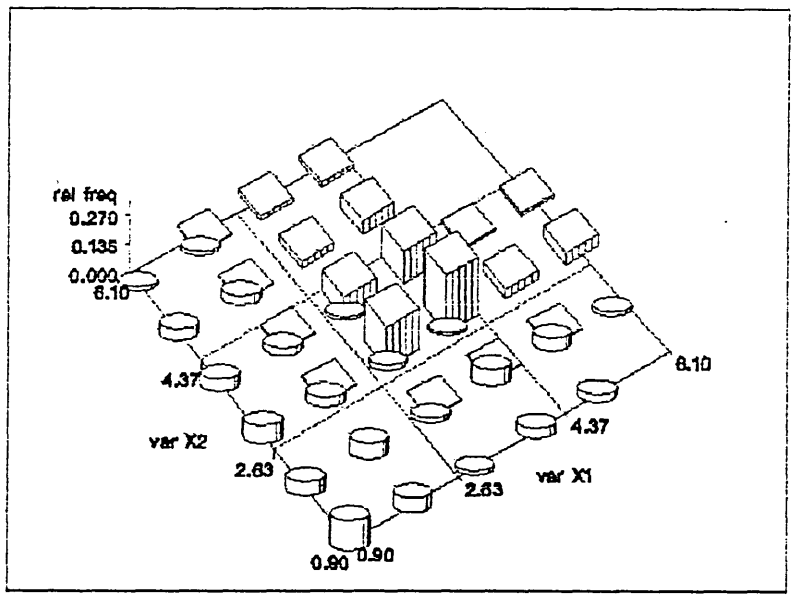


Figure 15.5-1: *3*-dimensional plot of *BANANA* data

Technical and theoretical admissibility: Due to the ordinal nature of the predictors the Bahadur models are inadmissible. Visual inspection shows that curvilinear separation lines are required for optimal discrimination. The *BANANA* dataset consists of "close to continuous" data with clearly different covariance matrices. This is the classical "text book" situation for the quadratic discriminant function. One might also expect the linear discriminant to do moderately well in terms of precision and bias. The distance based procedure by contrast would not be expected to perform well because no clear centroids can be established.

Due to the obvious patterned structure of the data there is a good case for applying the indirect recursive partitioning or the neural network approaches. As there are no symmetric dispersions of the data for both populations the euclidean distance based centroid procedure is ruled

out. The close to continuous metric implies that kernel density estimation based procedures should do well as they attempt to obtain local estimates of the underlying density. The data finally show considerable separation of the populations and moderately few cells considering the sample size which suggests that the multinomial model similarly should perform well.

Initial choice of procedure: Quadratic discriminant procedure or recursive partitioning procedure.

Density estimation: As specified by the *QDF*.

Crossvalidation of performance criteria: The medium to large sample size points to the hold-out crossvalidation technique.

Assessment: The results of table 15.5-1 summarised from the appendix confirm the expectation that the linear discriminant is not suitable for data with such strongly marked patterns. Within the set of available procedures[92] the picture is heterogeneous. This may suggest that neither procedure is optimally suited.

Inspection of the non-thresholded performance values, however, reveals that with $\varepsilon^{counting} = 0.036$ for the *MLT*, $\varepsilon^{posterior} = 0.033$ for the *DD2* and $\eta = 0.964$ for the *DHL* procedure, respectable separation may be achieved using discriminant procedures not initially selected.

statistic	$\varepsilon^{counting}$	$\varepsilon^{posterior}$	η
expectation	*MLT*	*DD2*	*DHL*
uncond s.e.	*MLT/DD2/DHL*	*KER*	*KER/DHL*
uncond bias	*KER/MLT*	*KER*	*DHL*

Table 15.5-1: Procedure choice for *BANANA* data

[92] the quadratic discriminant function as well as *CART* and *FACT* were not programmed.

Recommendations for final choice: Considering the above good performance of procedures not initially chosen on theoretical and empirical grounds one might settle for one of these. However, choice of the recursive partitioning procedure or the quadratic discriminant would be expected to produce optimal results.

15.6 Procedure choice for the *MA4353..* datasets

The *MA435300* to *MA435309* series of artificial datasets was constructed in order to inspect the behaviour of interaction models for dichotomous data. It consists of *10* samples (*MA435300* to *MA435309*) with *4* dichotomous predictors and equal priors $\pi_i = 0.500$. The data were generated according to the Bahadur model such that the data exhibit interactions between *2* variables and also between *3* variables taken at a time. Thus higher order interactions are deliberately built into the data.

Technical and theoretical admissibility: On theoretical grounds the corresponding *2*nd order and *3*rd order Bahadur are obvious candidates. Technically all procedures are admissible due to the dichotomous predictors. As with the *CESAR4* data there is a possible case for the centroid procedure and also possibly for the linear discriminant. However, to the extent that the modelled interactions are strong enough this would count against these two procedures. Empirical analysis of the data characteristics in chapter 11 shows that the artificial data exhibit the features of the *3*rd order Bahadur model with strong and significant correlations between all pairs of variables. As specified the correlations are all higher in Π_1 than in Π_2. The relative differences between the mean vectors are also evident. This is confirmed by the significant main effects for X_1 to X_4 from the loglinear analysis. The absence of significant interactive effects, however, is surprising.

Initial choice of procedure: One of the Bahadur models.

Density estimation: This follows directly from the specifications of the respective direct procedures.

Crossvalidation of performance criteria: The datasets are small, so leaving-one-out crossvalidation is recommended.

Assessment: In order to assess the optimal procedure from the results for the datasets *MA435300, ..., MA435309* individual values of performance criteria had to be averaged over the *10* samples. The following table gives averages of performance estimates for different procedures extracted from appendix A.

	mean values of		
proce-dure	$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
BH1	*.419*	*.261*	*.660*
BH2	*.418*	*.262*	*.660*
BH3	*.419*	*.261*	*.660*
KER	*.363*	*.311*	*.663*
LDF	*.402*	*.342*	*.628*
LG1	*.487*	*.379*	*.567*
MLT	*.353*	*.266*	*.690*
CEN	*.421*	*.262*	*.589*
DD1	*.292*	*.266*	*.706*
DHL	*.352*	*.264*	*.691*

Table 15.6-1: Average expected performance.

The results from the above table for expected conditional hold-out based estimates as well as corresponding estimates of unconditional standard error and absolute unconditional bias from appendices C and E, respectively, were condensed. In all cases averages were computed across the series of *10* datasets from corresponding values tabulated in the appendix. The procedure (or set of procedures) yielding best values was then entered into table 15.6-2[93]. The poor results for all Bahadur procedures in terms of non-

[93] for instance the *DD1* procedure had the lowest $\varepsilon_{counting}$ value of *0.292* in the previous table. N.B.: the conditions for entry are optimal values taken at an accuracy well beyond *3* significant post decimal digits.

thresholded performance as measured by $\varepsilon^{counting}$ and η was not expected on theoretical grounds. Only the logistic and the centroid procedure yield worse performance. By contrast the posterior based error rates, $\varepsilon^{posterior}$, are smallest for the Bahadur procedures. The *BH2* and *BH3* procedures also show least absolute bias for $\varepsilon^{counting}$ and η. The fact that the performance estimates for the Bahadur procedures in table 15.6-1 are all fairly similar, when compared with those for other procedures, is seen to indicate that the data may reveal an even more complex interactive structure.

statistic	$\varepsilon^{counting}$	$\varepsilon^{posterior}$	η
expectation	*DD1*	*BH1/BH3*	*DD1*
uncond s.e.	*BH1*	*KER*	*KER*
uncond bias	*BH2*	*LDF*	*BH3*

Table 15.6-2: Procedure choice for *MA4353* data

Recommendations for final choice: Depending on particular demands placed on the procedure a repetition of the cycle may be indicated because of the heterogeneous results above. They must be taken as evidence that caution is required when using the class of Bahadur procedures. It is possible that the data contain a 4th order effect that hasn't been modelled. Use of a corresponding *BH4* procedure, however, bears the danger of a high number of parameters. Looking again at table 15.6-1 the results of non-thresholded performance in terms of $\varepsilon^{counting}$ and η for the *DD1* procedure are convincing so that this nonparametric indirect method is recommended as a useful alternative procedure when faced with data exhibiting strong interactions such as the *MA4353* series.

15.7 Procedure choice for the *INTERAC1* dataset

The artificial *INTERAC1* dataset was constructed to inspect performance of a discriminant rule for datasets showing strong interactive effects.

Technical and theoretical admissibility: Visual inspection reveals that objects belonging to both populations are concentrated at two centroids. Optimal discrimination with single separation lines is therefore not possible. This is a classical case for recursive partitioning. In recursive partitioning such as *CART* and *FACT* the sample is split repeatedly, yielding a set of subsamples that are then allocated to the most likely parent population commonly assessed on a majority principle. The basic assumptions rule out the linear discriminant. The evident interaction may be modelled by a second order logistic. Both variables X_1 and X_2 are ordinal - $X_1 \in \{1,2,3,4\}$, $X_2 \in \{2,3,4,5,6,7\}$ - and so lead to comparatively smooth distributions. Thus again technically there is a case for the *LDF*, *QDF* and kernel procedures. But as seen above the shape of the data completely rules out the procedures based on single separation lines. The centroid is also ruled out because data distributions for both populations are multimodal. The distributional distance and *DHL* procedures operate directly on the vector of state probabilities and should not be affected by this feature of the data. Considering the total sample size the number of cells is moderate, which also speaks for the multinomial model where under these conditions the individual cell estimates will be fairly stable.

Initial choice of procedure: A recursive partitioning procedure and next any distance based procedures other than the centroid.

Density estimation: Not required for indirect procedures.

Crossvalidation of performance criteria: The dataset is very large, so one could possibly even accept resubstitution estimates as satisfactory.

Assessment: Table 15.7-1 shows results for non-threshold performance summarised from the appendix. Inspection of all non-thresholded performance criteria from tables A-1, A-2

and A-3 in the appendix confirms expectations on theoretical and empirical grounds in that certain procedures do extremely well while others fail completely. The comparatively good performance for the *DD1* ($\varepsilon^{counting} = 0.069$), *CEN* ($\varepsilon^{posterior} = 0.066$) and the *DD1*, *DD2* and *MLT* ($\eta = 0.932$) procedures suggest these as useful alternatives.

statistic	$\varepsilon^{counting}$	$\varepsilon^{posterior}$	η
expectation	*DD1*	*CEN*	*DD1/DD2/MLT*
uncond s.e.	-[94]	*DD1*	*DD1/MLT*
uncond bias	*DD2*	*MLT/KER*	*DD2*

Table 15.7-1: Procedure choice for *INTERAC1* data

Recommendations for final choice: As was the case with the *BANANA* data on theoretical and empirical grounds optimal performance is to be expected from recursive partitioning based procedures. However, the above distance based procedures show acceptable performance and could be used as efficient alternatives. Considering the dangers inherent in overfitting of *CART* or *FACT* procedures (see chapter 6) it might well be advisable to use a *DD1* procedure for datasets exhibiting strong interactions and patterns. If low error rates are not quite as important as reliable predictors then the high value of *0.932* for η also suggests the *MLT* procedure which is far easier to apply.

15.8 Procedure choice for the *IRIS* dataset

A modified version of the *IRIS* real dataset (Fisher's (1936) classical iris flower data) was chosen to inspect how the linear discriminant copes with discretised, originally normally distributed data. The data consists of separately sampled petal and sepal leaves of irises

[94] Estimates for the unconditional standard errors and unconditional biases are so similar that no clear preference for any one procedure is evident.

measured in terms of width and length for *3* populations (*virginica, setosa* and *versicolor* with equal priors π_i.

Technical and theoretical admissibility: In a sense the original data may be thought of as "designed" for applications of the linear discriminant function. It is reasonable to assume that even after the extreme transformation of chopping the distributions at interquartile percentage points, thus producing *3*-level ordinal variables, some of the former suitability should have remained. Therefore the linear discriminant will still be expected to produce satisfactory results. Separation of the populations should be fairly good as inspection of univariate relative differences in means shows (chapter 11). The original data exhibited quite strong correlations between dimensions of width and length of leaves. Even after discretisation this is still present as may be seen from inspecting the correlation matrices in chapter 11. This is also emphasised by the significant interactive effect for X_1 and X_2 after conducting a loglinear analysis of the data. The reduction to comparatively few discrete cells due to the interquartile range splits leads to stable estimates and thus the multinomial model is a further option. As the underlying densities are normal there is also a good case for the kernel procedure.

Initial choice of procedure: The linear discriminant function.

Density estimation: As specified in chapter 4.

Crossvalidation of performance criteria: Medium to small such that leaving-v-out or leaving-one-out crossvalidation is required.

Assessment: Table 15.8-1 shows results of non-thresholded performance summarised from appendix A, C and E. The observed estimates generally confirm expectations. The *LDF*

yielded expected performances of *0.128*, *0.052* and *0.910* for $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η, respectively.

The corresponding analysis of thresholded performance curves from chapter 14 is also in good agreement with choice of the linear discriminant procedure.

statistic	$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
expectation	*LDF*	*LDF*	*LDF*
uncond s.e.	*LDF*	*KER*	*KER*
uncond bias	*LDF*	*KER*	*LDF*

Table 15.8-1: Procedure choice for *IRIS* data

Recommendations for final choice: The linear discriminant procedure because of the unequivocal results. However, if precision is more important than a low error rate then the kernel density estimation based procedure ought to be considered as a second option.

15.9 Procedure choice for the *EDUC* dataset

The *EDUC* real dataset was chosen because of its multi-level ordinal response and ordinal predictors. It is a comparatively large dataset with priors $\pi_1 = 0.192$, $\pi_2 = 0.066$, $\pi_3 = 0.054$ and $\pi_4 = 0.689$. The dependent occupational group recorded at *4* ordinal levels is related to scholastic ability rated at *5* ordinal levels and educational achievement rated at *4* ordinal levels.

Technical and theoretical admissibility: The Bahadur procedures are excluded on technical grounds. The ordinal response calls for discriminant procedures aimed at modelling this feature. Here the cumulative logit model is an obvious candidate. Other procedures are linear regression or regression trees as used in *CART* or *FACT*. The ordinal nature of predictors also suggests entering the

predictors directly into the logistic model rather than as class variables - using the terminology of *SAS*. These will be strong as achieved educational level is related to income and so education will be correlated with ability. This indicates on theoretical grounds a procedure that models the interactive data structure, such as a higher order logistic. The fact that there are *4* and *5* levels respectively for each of the predictors puts the data into a class where the ordinal nature approaches continuous structures and thus it is to be expected that procedures for continuous distributions such as the *LDF* may produce adequate results as well. Even though the number of states is fairly large with $s = 20$ the sample sizes per population are sufficiently big to allow for stable parameter estimates in the multinomial procedure.

Consideration of the univariate statistics from chapter 11 shows the mean vectors to be fairly similar across all populations. The correlations between X_1 and X_2 are significant, though small in the two larger populations Π_1 and Π_2. Loglinear analysis reveals significant main effects. On the basis of the above empirical analysis therefore, there is less support for a procedure based on modelling of interactions and instead a stronger argument for a more parsimonious model such as in the multinomial procedure.

Initial choice of procedure: Cumulative logit based procedure, alternatively the *LDF* or the *MLT* procedure.

Density estimation: As specified by the chosen procedure.

Crossvalidation of performance criteria: Large dataset, so resubstitution estimates might suffice. The biases can be expected to be small overall.

Assessment: Table 15.9-1 shows results of non-thresholded performance summarised from appendix A, C and E.

statistic	$\varepsilon_{counting}$	$\varepsilon_{posterior}$	η
expectation	*LDF/KER/MLT*	*DD2*	*LDF*
uncond s.e.	*KER/LDF*	–[95]	*KER*
uncond bias	*KER*	*DD1*	*MLT*

Table 15.9-1: Procedure choice for *EDUC* data

The good performance of the distributional distance model (*DD2*) as measured by the posterior based error rate estimator $\varepsilon_{posterior}$ may be attributed to its ability to deal well with cell heterogeneity.

Expected values of performance criteria are *0.311*, *0.310* and *0.710* for $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η respectively. The *KER* procedure yielded an η value just slightly smaller than *0.710*. Considering the comparatively small priors for some of the populations this shows poor separation. These results must be taken to imply that the information contained in the predictors is not specific enough to enable successful discrimination.

Recommendations for final choice: For the particular problem addressed by this dataset it is highly recommended that in addition other predictors should be looked into. If this proves technically or otherwise impossible then the results produced by the linear discriminant or kernel procedure have to be accepted as best under the circumstances.

15.10 Conclusions

The selection tree developed in chapter 12 is used to choose optimal discriminant procedures for selected real and artificial datasets. The selection process consists of the stages:

[95] Estimates of the unconditional standard errors are so similar that no clear preference for any one procedure is apparent.

(1) analysis of theoretical and technical admissibility, taking potential information on the underlying data model as well as empirical information on the sample into consideration,

(2) initial choice of procedure on the basis of step (1),

(3) choice of density estimation technique (in the case of direct procedures),

(4) choice of crossvalidation technique,

(5) assessment of initial choice on the basis of hold-out based estimates of non-thresholded performance criteria and, where given, on the basis of thresholded performance, and finally

(6) optional repetition of the selection cycle given demands and constraints particular to the given discriminant problem.

Assessment is made in terms of expected conditional hold-out based values, expected unconditional standard errors and expected unconditional bias estimates. Both non-thresholded and thresholded results are compared for $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η.

Good agreement between initial choice by the selection tree and actually observed performance as given in the respective summary tables was observed for the *CESAR4*, *CREDIT* and the *IRIS* datasets. Initial choice confirmed by observed performance fell on the third order Bahadur and on the linear discriminant procedures for the *CESAR4* dataset, on the Hills distance and the modified distributional distance procedure for the *CREDIT* data and on the linear discriminant procedure for the *IRIS* data.

It was expected that the Bahadur procedures would perform better for the artificial *MA4353* series of datasets because they were deliberately constructed to include second and third order interactions. The finding that all Bahadur

procedures showed similarly poor performance is seen to indicate caution when using the class of Bahadur procedures. Good results were achieved using expected values of $\varepsilon^{posterior}$ while performance in terms of expected values of $\varepsilon^{counting}$ and η was distinctly better for other procedures. It is therefore concluded that the conditions that enable successful performance of the Bahadur procedures depend on exact modelling of the data structures which may require higher order models. These, however, bear the danger of overparametrisation.

In the case of the *CESAR4* dataset comparisons between thresholded performance (chapter 14) and non-thresholded performance showed that both approaches can lead to different procedure selections. This was particularly marked when performance is judged using $\varepsilon^{counting}$. Non-thresholded results suggest the *BH3* procedure, thresholded results point to either of the *KER*, *LDF*, *CEN* or *MLT* procedures. The fact that threshold dependent analysis suggests several procedures is not to be seen as a weakness of this approach as inspection of the corresponding performance curves for $\varepsilon^{counting}$ (see appendix G) shows. These lie close together for almost the entire range of τ. Set against this one would evidently draw false conclusions in singling out the *BH3* procedure as optimal on the basis of non-thresholded performance alone.

The analysis of the *EDUC* data revealed that the predictive power of the independent variables is poor and a recommendation for collecting further data must be made. In the case of the *CHD* data a recommendation was made for adoption of different costs of misallocation or alternatively using the selective discrimination approach.

Initial choices for the *BANANA* and *INTERAC1* datasets were not confirmed because the recursive partitioning based procedures were not implemented. On theoretical grounds their performance may be expected to be optimal.

Summarising the above it is concluded that application of the selection tree leads to clear end points in each case. These may be as varied as choice of a theoretically inappropriate procedure, choice of procedure *A* if precision is more important than low error rates and *B* otherwise or even choice of no procedure but instead recommendation for gathering new data. Key stages in the selection process are reached when a decision has to be made concerning (a) the use of theoretically inappropriate procedures (b) the relative importance of low error rates (or equivalently high η values) and precision in terms of bias and variance, (c) the use of additional information contained in η and (d) the interpretation of thresholded performance.

I: INTRODUCTION

II: REVIEW

III: METHOD

IV: RESULTS

V: DISCUSSION

16. Discussion 16.1 Literature review 16.2 Construction of performance criteria 16.3 Construction of selection trees
17. Further Studies
18. Conclusions

The research title is "The Selection of Optimal Discriminant Procedures for Discrete Data". Taking each of the key terms in turn "Selection" implies the *guided choice* and also the *decision tree*. "Optimal" implies that the choice is the *best among a set of all procedures considered* and also a choice that is *best under the given information about the data*. "Discriminant" implies some suitable *success criterion by which to judge a procedure's performance*. The reference to "Procedures" implies the discussion of *standard procedures, augmented procedures* and *entirely new procedures*. Finally the emphasis on "Discrete Data" leads to the discussion of *where continuous data becomes discrete*. All these ideas have been addressed in the thesis. The introductory chapter 2 gave some examples of typical situations in discriminant analysis and identified common problems faced by users of discriminant procedures. It was felt that these key issues are not adequately addressed in the literature. A need for further work was seen especially

(i) in the field of discrete data with a limited number of states, and

(ii) in particular concerning the provision of guides to optimal procedure selection.

In order to answer this need the aims of the research were thus stated at the end of chapter 2 as follows:

(a) review of the relevant literature,

(b) construction of performance criteria suitable for discrete data, and

(c) construction of a general (and formal) guide to procedure selection suitable for discrete data as opposed to continuous.

The first aim (a) is addressed in chapters 3 to 6, (b) in chapters 8 and 9 and (c) in chapter 12. The tackling of these aims is now discussed in separate sections.

16.1 Literature review

As expected the literature review gives only limited help in procedure selection. On the other hand there is evidently a considerable need for guides to optimal procedure selection. This became clear from the results of the *MEDLINE* literature search conducted for the publication years 1989, 1991 and 1993. The popular linear discriminant function was frequently used also in data situations characterised by marked departures from normality. Given more readily available structured guides to optimal procedure selection this situation might well change in future years.

Inappropriate choice of procedure, however, is in no way due to a lack of variety of discriminant procedures for discrete data. The literature is abundant with a breadth of different procedures. These range from the direct parametric and nonparametric procedures to the class of indirect procedures.

Recent work is predominantly in computer intensive areas usually involving iterative techniques such as in nonparametric density estimation, recursive partitioning or artificial neural networks. Although the availability of powerful computer resources is ever increasing the recent trend towards more machine intensive techniques does raise the question of whether the particular demands given in any concrete situation always justify such elaborate and expensive procedures.

The successful execution of some of these procedures such as the kernel density estimation based ones proves to be so sensitive to settings for the smoothing parameters that

this consequently opens up the field of nonparametric density estimation. This is discussed in chapter 5.

Success of a procedure will be judged in terms of the qualities of the characteristics expected from the procedure when applied to new data. These qualities are usually measured by means of suitable performance criteria such as the error rate or its expected bias. Thus it is evident that choice of performance criteria will precede the selection of a discriminant procedure.

The central role played by performance criteria in the process of procedure selection is made clear in section 4 of chapter 12 where different aspects of performance are summarised. These break down into (1) misallocation errors or hit rates, (2) separation measures, (3) bias reduction or crossvalidation methods and (4) reliability measures.

16.2 Performance criteria

Two new means of assessing performance based on the posterior distribution were constructed: the η criterion and the concept of variable classification thresholds.

In section 3 of chapter 8 it was noted that under certain conditions the information contained in the entire distribution of posteriors is at least a large as that in the subset of posteriors for correctly allocated objects. This was illustrated by hypothetical datasets and forms the theoretical basis for the construction of η. By contrast the posterior probability based error rate estimator, $\varepsilon_{posterior}$, of Hora and Wilcox (1982), is based only on the posteriors for correctly allocated objects.

The new posterior based η criterion generally exhibited lower variance and better bias characteristics than $\varepsilon_{counting}$ and frequently also than $\varepsilon_{posterior}$. An internal crossreference between the three performance criteria

indicates that their validity is given because of the generally high degree of correlation among all three[96].

Because the η criterion balances posteriors for correctly allocated objects as well as for misallocated objects this criterion is essentially different in nature to the error rate estimators. η is not seen as a substitute for estimators of the misallocation error but with respect to procedure selection as an additional criterion of performance. Its advantages lie in its low bias and variance characteristics and as such η provides a useful measure of reliable assessment of performance of a discriminant procedure.

The question as to whether the lower variance of η would mask differences in performance that other criteria might pick out was discussed in chapter 13. It was found that this was not the case. In several instances η revealed different values for different procedures applied to the same dataset, while the misallocation error $\varepsilon^{counting}$ showed identical or very similar values.

Variable classification thresholds are derived from relative differences between posteriors. The analysis of the distributions of relative posterior differences was originally not intended but emerged as a further diagnostic aid in the course of the research. Although it was felt at the very outset that the distribution of posterior probabilities held the key to more differentiated evaluation of a discriminant's performance, the plot of relative differences appeared indirectly. It came about by considering the effect of using a classification threshold. In particular the idea of variable classification thresholds led to the concept of relative differences between the two largest posteriors $\tau = (h^{(1)}-h^{(2)}) \; / \; h^{(1)}$.

96

As η is scaled oppositely to $\varepsilon^{counting}$ and $\varepsilon^{posterior}$ the correlation between the error rate estimators is positive while correlations with η are negative.

The interpretation of the estimated distributions of $f(\tau)$ in chapter 14 is not always straightforward. Generally it appears that an initial guess at a promising procedure is possible but this is not always so as in the case of the *NORMAL16* data example (chapter 14). Here the estimated $f(\tau)$ distribution suggests that nonparametric distance based procedures and the logistic procedure will perform well. However, as the performance curves later show - and this is also in line with the origin of the data (see chapter 11) - only the linear discriminant is singled out by the leaving-one-out technique for all three performance criteria in terms of absolute level. Hence one must conclude that the inspection of the estimated distributions of $f(\tau)$ only gives an indication and that the crucial insight must be gained from the performance curves.

As was pointed out at the end of chapter 14, what is not immediately obvious is that inspection of a procedure's performance φ over the whole range of relative posterior differences in $0 < \tau \leq 1$ can be very different from just using $\varphi|_{\tau=0}$. Non-thresholded performance is just the initial point of the performance curve ranging from *0* to *1*. Normally it is only the initial section of the range up to about *0.3* that will be of interest and the focus should be on the first part of the curves when judging performance. The reasons for this are twofold. From the analysis in chapter 14 it was seen that performance curves spread out rapidly at low thresholds. In practical applications furthermore only moderate classification thresholds will be used to avoid too many rejections. It is not uncommon that the estimates of non-thresholded performance, $\varphi|_{\tau=0}$, give a different ranking of procedures than would obtain by rating performance in terms of the curves.

The performance curves may initially lie close together thus making it difficult to select any one procedure in preference of another. However, once they have separated they do show considerable stability over τ with few reversals of ranking thus making choice of procedure a

comparatively easy task. Because performance curves have a tendency to spread out as the threshold τ increases, an entire group of procedures can often be quickly eliminated from further consideration. This feature of thresholding performance is seen as particularly useful when reliability of a procedure is important.

In summary, the advantage of thresholding allocations in discriminant analysis lies in the greater degree of separation achievable. The approach gives greater reliability of the estimated rule in the presence of sampling variations. Thresholding should however be restricted to the lower end of the τ range because, although desirable, large differences in the two largest posteriors τ will also lead to higher rejection rates. As a consequence focusing on the lower end of the τ scale below about *0.30* is suggested. The analyses revealed that generally performance declines rapidly at higher thresholds. It is further also possible to combine the benefits of the posterior based η criterion, as well as any of the other criteria, with the advantages of classification thresholds.

16.3 Selection trees

As pointed out in section 16.1 the review showed that an abundance of discriminant procedures for discrete data exists, yet few structured guides to selection are available. It was decided at an early stage that greater benefit would be gained from developing general guidelines for approaching the selection problem rather than providing a catalogue of all possible discriminant procedures. From the outset it was therefore clear that a selection strategy would have to be developed on a subset of all procedures. As stated at the end of chapter 2 the intention was not to present a *cookbook* of procedures but rather to identify the critical factors leading to optimal choice.

Five key factors influencing choice were identified: *model* information, *data* information, *demands* placed on the discriminant, *constraints* and *skill* or *experience* of the user. Of these the first three are formally placed in the suggested selection tree. The other two factors were excluded as they are respectively either subject to developments in the hardware and software industry or difficult to quantify. Due to the wide range of possible performance criteria the selection tree allows for an iterative process of identification of optimal procedures.

The traditional approach to procedure selection is the classical one where it is assumed that sufficient information about the statistical model is available to enable unequivocal choice of procedure. This is termed selection in terms of *theoretical admissibility*. Other more pragmatic approaches begin by starting with all procedures that can be technically applied to the data and then narrowing down choice for instance by minimising the error rate. These approaches are termed selection in terms of *technical admissibility*.

The proposed selection tree combines both approaches above by considering *model* and *data* information upon which an initial selection is made. In situations where the model information outweighs data information selection will be classical. When there is more data information selection will be more pragmatic. Initial selection will thus reflect a balance between classical and pragmatic approaches.

Different performance criteria measure different qualities of performance. By offering a range of performance criteria the selection tree can be adjusted to different demands. This involves deciding among $\varepsilon_{\text{counting}}$, $\varepsilon_{\text{posterior}}$ and η as well as the analysis of threshold dependent performance. If low variance is a priority one might select η. If considerable sampling variation is to be expected, perhaps because of small sample sizes, then threshold dependent performance curves plotted against τ should be selected.

The conclusion to be drawn from the worked examples is that as a rule of thumb inspection of the behaviour of a discriminant at the lower end of the threshold spectrum ought to be carried out if possible in situations where non-thresholded performance values lie close together.

After execution of the initially selected procedure analysis of performance as judged by the chosen performance criteria will indicate whether the selection process has to be repeated for another cycle. Use of the selection tree in the above manner may result in the selection of procedures that do not always match underlying data models. Under certain conditions - frequently when only small datasets are available - this leads to locally optimal, yet theoretically inappropriate procedures. This results does not come as a surprise as Victor (1976) demonstrated that often theoretically inappropriate yet technically admissible models can lead to acceptable results.

I: INTRODUCTION

II: REVIEW

III: METHOD

IV: RESULTS

V: DISCUSSION

16. Discussion

17. Further Studies

17.1 Data

17.2 Procedures

17.3 Performance criteria

18. Conclusions

Even in an extensive treatment of a given subject such as the present some topics will prove to be so loosely related to the central theme that they have to become side issues perhaps to be pursued in the context of other research work. The following sections briefly list the topics that were not dealt with in depth. In each case the consequences of ignoring these aspects are briefly discussed to allow an appraisal of relevance.

17.1 Data

When contrasted with continuous data in the classical sense some of the datasets analysed may strike one as being "too discrete". However, a treatment of discrete discriminant analysis that stops short of an in depth consideration of datasets that are essentially variants of contingency tables fails to meet the common demands for discriminant problems in the social sciences, in market research or medicine. This was made clear in section 1 of chapter 3. Frequently as a consequence the number of cells in the datasets analysed is comparatively small.

The number of datasets presented in chapter 14 as examples of analysing threshold dependent performance might be increased. It can be expected, however, that this will not change the overall conclusion reached in chapters 14 and 15, namely that use of posterior probability based performance criteria and especially plots of threshold dependent performance help to assess the reliability of discriminant procedures.

If other artificial datasets were to be explored it might be useful to include further data showing different patterns of covariances between the predictors. This could shed further light on the failure of the Bahadur procedure to perform well for the *MA4353* series of datasets.

The more recent recursive partitioning and artificial neural network procedures were mentioned in the context of discriminant analysis. It was decided to treat these theoretically, and instead to focus experimental work on other procedures. It could be argued that to complete the picture actual worked examples possibly also for higher order logistic, cumulative logistic and quadratic discriminants are required. While this seems inherently plausible it must be stressed again that the chief aim of the research was to develop a methodology by which selection from a range of procedures could be achieved in a systematic fashion.

Where applicable, and especially in the worked selection examples of chapters 14 and 15, other procedures that appeared promising in the given case were referred to and critically discussed as candidates for satisfactory performance. Regarding the artificial neural networks some theoretical assessment was made in chapter 6. Here the general conclusion was reached that for the majority of typical discriminant analysis problems in social science, market research or medicine procedures based on *ANN*'s do not clearly appear promising because of the equivocal results reported in the literature and especially because of few reported applications to discrete data. Classical areas of application for *ANN*'s are still in pattern recognition and image analysis such as used for military purposes or in astronomy. These typically involve continuous data.

An extension of the kernel discriminant procedure to include the optimisation of the bandwidth parameter λ[97] could help to better illustrate the wide applicability of this nonparametric procedure. For the present purposes of demonstrating the philosophy behind procedure selection it

[97] see chapter 5

is not required. However, when seriously considering a kernel based discriminant procedure the bandwidth parameter λ should be optimised. This is commonly done using bootstrap techniques.

The results presented in chapters 14 and 15 frequently revealed better performance for the Bahadur models and for the population distance based procedures such as *DD1*, *DD2* and *DHL*. This was, as expected, mostly the case when the dataset exhibited non-normal structure. In the light of this it might be worth considering the inclusion of these models in statistical packages such that discrete datasets may be analysed more flexibly.

17.3 Performance criteria

In the case of datasets with dichotomous outcomes it may be argued that estimates of sensitivity and specificity such as used in receiver operating characteristic curves (*ROC*) could help to display the performance characteristics of chosen discriminant procedures. In classical linear discriminant analysis *ROC* curves may be obtained when the cutoff point is shifted laterally along the line intersecting both population centroids. Normally, when the priors are known or can be estimated, an optimal cutoff point is found by correcting the linear discriminant function by the logarithm of the ratio of the priors, $\log_e(\pi_2/\pi_1)$. For any cutoff point one sensitivity estimate and one specificity estimate result.

If there is doubt about the validity of the estimated priors $\hat{\pi}_i$ then interest may lie in exploring the performance of discriminant rule in a region R around the prior estimate $\{\hat{\pi}_i \pm b\}$. This essentially Bayesian approach[98] has been investigated empirically by inspecting the performance of discriminant procedures over a range of

[98] see chapter 4

prior probabilities in the context of an MSc project (Lack, 1987).

In the present thesis the question of stability of estimates is addressed with respect to the variable classification threshold, τ. These results are presented in chapter 14. Here it may be argued that the estimates of performance criteria plotted against τ should be augmented by corresponding estimates of bias and standard error as well.

Further an empirical inspection of the behaviour of the formal performance criterion, φ^*, suggested in chapter 9 might help to corroborate the expectation that it is highly correlated with φ^{99}.

The discussion of performance curves showed that interest would focus especially on values of τ near the origin. It is not always clear from the plots how the procedures are to be ranked at the lower end of the τ range where marked changes in values of $\varphi_{|\tau}$ can occur. Here it might be of help to magnify this portion of the scale appropriately.

A final extension to the analysis of performance criteria might also include further correlation analyses between customary error rates and alternative posterior probability based criteria as well as additional analyses of the threshold dependent performance presented in chapter 15.

99

where φ stands for $\varepsilon^{counting}$, $\varepsilon^{posterior}$ or η

I: INTRODUCTION

II: REVIEW

III: METHOD

IV: RESULTS

V: DISCUSSION

16. Discussion

17. Further Studies

18. Conclusions

18.1 Performance criteria

18.2 Classification thresholds

18.3 Identification of key factors of choice

18.4 Construction of a procedure selection tree

18.5 Validity of selection tree

18.6 Extended notation for performance criteria

18.7 Ease of implementation

18.8 Use of datasets

18.9 Discriminant procedures

The non-technical examples given in chapter 2 outline the general problem of discriminant analysis for discrete data. The subsequent literature review presented in chapter 3 reveals that few if any general guides exist, particularly with applications to discrete data. Although there are several books on discriminant analysis for discrete data they generally focus on description of procedures and leave the choice largely to the user.

The need for suitable guides to aid the researcher wishing to employ discriminant techniques is made apparent by a recently conducted literature search of the *MEDLINE* data base for the publication years 1989, 1991 and 1993[100]. In medical research considerable use still appears to be made of the linear discriminant function even in data situations where departures are so far from normality that other more appropriate discriminant procedures would show clear advantages in terms of performance.

The above state of affairs motivated the research for establishing suitable principles of discriminant procedure selection for discrete data. The statistical tool of discriminant analysis is not as narrowly defined as linear regression, for instance. Discriminant analysis embraces a wide variety of individual component techniques such as sampling, model building, density estimation, error rate analysis and reliability assessment. Any attempt to address the problem of selection of discriminant procedure must therefore include an appraisal of existing techniques related to discriminant analysis. For this reason the review embraces the topic of *nonparametric density estimation* (chapter 5) which is particularly relevant to discrete data and the topic of *performance evaluation* (chapter 7) as well as a thorough treatment of *direct* and *indirect procedures*. The methodology developed for

[100] see chapter 2 for details

performance criteria (chapter 8), *classification thresholds* (chapter 9) as well as *specific adjustments to procedures and crossvalidation techniques* (chapter 10) leads to the *construction of selection rules* (chapter 12) that are applied to a range of real and artificial datasets (chapter 11).

Was it possible to find such suitable principles for "The Selection of Optimal Discriminant Procedures for Discrete Data" ? The results from the above research (chapters 13, 14 and 15) as well as the discussion in chapter 16 allow the following conclusions to be drawn.

18.1 Performance criteria

It was observed (chapter 8) that under certain conditions the information contained in the entire distribution of posterior probabilities across discrete data states may exceed that contained in the subset of correctly allocated objects commonly used in customary error rates. On the basis of this finding a new measure of performance of a discriminant procedure for discrete data was constructed. A major aim was to reduce the variance in customary error rates caused by sampling from discrete distributions. The posterior error rate estimator, $\varepsilon^{posterior}$, of Hora and Wilcox (1982) did not quite match up to the expectations of lower variance when compared with the customary counting based error rate, $\varepsilon^{counting}$. However, the suggested new eta criterion, η, did generally show lower variance when compared with $\varepsilon^{counting}$. The bias characteristics of posterior probability based performance estimators are overall satisfactory. Generally variance and bias are lowest for η. The lower variance of η when compared to $\varepsilon^{counting}$ does not lead to masking of differences in performance detectable by η. Empirical evidence suggests that η measures different qualities of a discriminant's performance. It is recommended to use η as an additional measure of non-thresholded performance especially when

$\varepsilon_{counting}$ yield similar values for different discriminant procedures.

18.2 Classification thresholds

Use of the distribution of posterior probabilities of population membership in the construction of η lead to a closer inspection of the distribution of relative differences $f(\tau)$ between the two largest posteriors. The idea of a *relative difference* is an extension of the posterior thresholding concept implemented in some statistical software packages for discriminant analysis. By considering the behaviour of performance criteria over the entire range of relative differences, $0 \leq \tau \leq 1$, further insight can be gained into the expected performance of a discriminant rule. Generally it is sufficient to look at the lower end of the τ-range because the respective performance curves soon spread out and tend to remain stable with respect to their rank order. In actual applications of posterior thresholding it is recommended to select a discriminant based on its performance within the lower third of the τ-range. Thresholded performance analysis is particularly useful when non-thresholded performance estimates for different procedures lie close together. Use of variable as opposed to fixed classification thresholds should be used for discrimination between three and more populations.

18.3 Identification of main factors of choice

Discussion of factors influencing selection of a discriminant procedure lead to identification of five critical factors: (1) theoretical information about the underlying data distribution, (2) empirical information on the data such as scaling of variables and sample size, (3) demands placed on expected performance of the discriminant rule, (4) availability of computer resources and finally

(5) skill and experience of the user. Key stages in the selection process are reached when a decision has to be made concerning (a) the use of theoretically inappropriate procedures (b) the relative importance of low error rates (or equivalently high η values) and precision in terms of bias and variance, (c) the use of additional information contained in η and (d) the interpretation of thresholded performance.

18.4 Construction of a procedure selection tree

A formalised approach focusing on the first three key factors identified above is developed resulting in a selection tree. The main feature of the tree is initial choice based on factors (1) and (2) and subsequent modification of initial choice after assessment of performance with respect to factor (3). Several applications of this selection tree to real and artificial datasets with different data structures are demonstrated.

18.5 Validity of selection tree

The initial choices derived from application of the selection tree were compared with detailed results for estimated performance. The comparisons do not necessarily confirm initial choice of discriminant procedure. Application of the selection tree, however, does lead to practicable recommendations. A major requirement for successful application of the selection tree is to fix the demands placed on a discriminant procedure. For instance, choice of procedure depends critically on whether non-thresholded or thresholded performance is used for assessment.

18.6 Extended notation for performance criteria

Selection of a discriminant procedure on theoretical grounds alone is no guarantee for good performance, as was pointed out by Christl and Stock (1973). The best example is the wide applicability of the linear discriminant because of its robustness. An illustration of these paradoxical findings was designed along the lines of a similar graph due to Victor (1976). In order to distinguish good performance of a discriminant procedure *due* to theoretical applicability from good performance *in spite* of missing theoretical applicability an extension of the common notation for the error rate was developed in chapter 7. This notation proves useful in the interpretation of results.

18.7 Ease of implementation

Application of the selection tree developed above is straightforward. The computation of η requires the posterior probabilities for each observation. Professional statistical packages generally supply these probabilities for standard procedures. The results, however, do show that less common procedures such as the Bahadur may occasionally be more appropriate. A need is seen for the integration of other discriminant procedures especially for discrete datasets. The variable classification concept should similarly be incorporated.

18.8 Use of datasets

A variety of different real and artificial datasets was used to demonstrate the response of the selection tree to different data structures and also to inspect the behaviour of the discriminant procedures when subjected to non-standard assumptions. As these datasets were used for illustration actual applications of the selection tree must

involve consideration of the specific characteristics of the given dataset and demands placed on the particular discriminant problem.

18.9 Discriminant procedures

The procedures used for comparative analyses were taken from the literature and constitute a subset of all procedures. In the case of Hills' (1966) distance based procedure a modification had to be carried out to allow extension to *3* and more populations. It was not essential to cover the entire range of discriminant procedures for developing guidelines for procedure selection. For the sake of completeness, however, some other popular procedures for discrete data such as variants of the kernel density or nearest neighbour based procedures, higher order logistic procedures, recursive partitioning based procedures and particularly the recent neural network based procedures are also discussed. With respect to the latter it is noted that presently few references to applications of discrete data exist. Some of the reported superior performances of artificial neural network procedures when compared to standard applications of discriminant analysis for continuous data need not necessarily imply superior performance for discrete data as well.

Adegboye, O.S. (1987). "A classification rule whose probability of misclassification is independent of the variance", *Australian Journal of Statistics*, 29, 208-213

Aeberhard, S., Coomans, D. and de Vel, O. (1994). "Comparative analysis of statistical pattern recognition methods in high dimensional settings", *Pattern Recognition*, 27(8), 1065-1077

Aitchison, J. and Aitken, C.G.G. (1976). "Multivariate binary discrimination by the kernel method", *Biometrika*, 63(3), 413-420

Aitchison, J. and Dunsmore, I.R. (1975). *Statistical prediction analysis*. Cambridge Univ Press: Cambridge.

Aitken, C.G.G. (1983). "Kernel methods for the estimation of discrete distributions", *Journal of Statistical Computing and Simulation*, 16, 189-200

Albert, A. and Lesaffre, E. (1986). "Multiple group logistic discrimination", *Computing and Mathematics with Applications*, 12A(2), 209-224

Allen, G. and Le Marshall, J.F. (1994). "An evaluation of neural networks and discriminant analysis methods for application in operational rain forecasting", *Australian Meteorological Magazine*, 43(1), 17-28

Amemiya, T.A. (1983). "A comparison of the logit model and normal discriminant analysis when the independent variables are binary", *Studies in Econometrics, Time Series & Multivariate Analysis*, (Karlin S), 3-30

Anderson, E. (1957). "A semi-graphical method for the analysis of complex problems", *Proceedings of the National Academy of Sciences, U.S.A.*, 43, 923-927, <reprinted in *Technometrics* (1960), 2, 387-392.>

Anderson, J.A., Whaley, K., Williamson, J. and Buchanan, W.W. (1972). "A statistical aid to the diagnosis of *keratoconjunctivitis sicca*", *Quarterly Journal of Medicine, New Series, XLI*, 162, 175-189

Anderson, J.A. (1975). "Quadratic logistic discrimination", *Biometrika*, 62(1), 149-154

Anderson, J.A. (1982). "Logistic discrimination", In *Handbook of Statistics*, 2, Krishnaiah, P.R. and Kanal, L.N., 169-191

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley

Andrews, D.F. (1972). "Plots of high-dimensional data", *Biometrics*, 28, 125-136

Arminger, G. and Sobel, M.E. (1990). "Pseudo-maximum likelihood estimation of mean and covariance-structures with missing data", *Journal of the American Statistical Association*, 85(409), 195-203

Asoh, H. and Otsu, N. (1990). "An application of nonlinear discriminant analysis by multilayer neural networks", *IEEE*, 3, 211-216

Bahadur, R.R. (1961a). "A representation of the joint distribution of responses to n dichotomous items", In *Studies in Item Analysis*, Solomon, H.(Ed.), Stanford: Stanford University Press, 158-168

Bahadur, R.R. (1961b). "On classification based upon responses to n dichotomous items", In *Studies in Item Analysis*, Solomon, H. (Ed.), Stanford: Stanford University Press, 169-176

Baichun Xiao (1994). "Necessary and sufficient conditions of unacceptable solutions in NLP discriminant analysis", *European Journal of Operational Research*, 78(3), 404-412

Balakrishnan, N. (1988). "Robustness of the double discriminant function in nonnormal situations", *South African Statistical Journal*, 22, 15-43

Ball, G.H. and Hall, D.J. (1970). "Some implications of interactive graphic computer systems, for data analysis and statistics", *Technometrics*, 12, 17-31

Baron A.E. (1991). "Misclassification among methods used for multiple group discrimination - the effects of distributional properties", *Statistics in Medicine*, 10(5), 757-766

Basford, K.E. and McLachlan, G.J. (1985). "Estimation of allocation rates in a cluster analysis context", *Journal of the American Statistical Association*, 80, 286-293

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The new S language*. Belmont, California: Wadsworth

Ben-Bassat, M. (1977). "Properties and convergence of a posteriori probabilities in classification", *Pattern Recognition*, 9, 99-107

Berres, M. (1993). "A comparison of some linear and nonlinear discrimination methods", *Computational Statistics*, 8(3), 223-239

Birch, M.W. (1963). "Maximum likelihood in three way contingency tables", *Journal of the Royal Statistical Society Series B*, 25, 220-233

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.

Bowman, A.W. (1984). "Cross-validation in nonparametric estimation of probabilities and probability", *Biometrika*, 71, 341-351

Bowman, A.W. (1984). "An alternative method of cross validation for the smoothing of density estimates", *Biometrika*, 71, 353-360

Bowman, A.W. (1985). "A comparative study of some kernel-based nonparametric density estimators", *Journal of Statistical Computing and Simulation*, 21, 313-327

Breiman, L., Olshen, R.A., Friedmann, J.H. and Stone, C.S. (1984). *Classification and regression trees (CART)*. Wadsworth, Belmont, California

Breiman, L. (1988). Contribution to the discussion of paper by W.H. Loh and N. Vanichsetakul, *Journal of the American Statistical Association*, 83, 725-727

Brigatti, L., Filatov, V., Hoffman, D., Assad, A. and Caprioli, J. (1993). "Comparison of Neural Networks with Discriminant-Analysis to Classify Glaucomatous and Normal Eyes", *Investigative Ophthalmology and Visual Science*, 34(4), 763-763

Broffit, J.D. (1982). "Nonparametric classification", In *Handbook of Statistics*, Vol 2 (Krishnaiah, P.R. and Kanal, L.N.), 139-168

Bruckner, L.A. (1978). *On Chernoff faces*. In *Graphical Representation of Multivariate Data*, Wang, P.C.C. (Ed.), New York: Academic Press, 93-121

Bryant, J. (1989). "A fast classifier for image data", *Pattern Recognition*, 22, 45-48

Bull, S.B. and Donner, A. (1987). "The efficiency of multinomial logistic regression compared with multiple group discriminant analysis", *Journal of the American Statistical Association*, 82(400), 1118-1122

Butler, W.J. and Kronmal, R.A. (1985). "Discrimination with Polychotomous Predictor Variables Using Orthogonal Functions", *Journal of the American Statistical Association*, 80(390), 443-448

Byth, K. (1980). "Logistic regression compared to normal discrimination for non-normal populations", *Australian Journal of Statistics*, 22, 188-196

Cacoullos, T. (1966). "Estimation of a multivariate density", *Annals of the Institute of Statistical Mathematics*, 18, 179-189

Campbell, M.N. and Donner, A. (1989). "Classification efficiency of multinomial logistic regression relative to ordinal logistic regression", *Journal of the American Statistical Association*, 84, 587-591

Campbell, N.A. and Mahon, R.J. (1974). "A multivariate sudy of variation in two species of of rock crab of genus leptograpsus", *Australian Journal of Zoology*, 22, 417-425

Čencov, N.N. (1962). "Evaluation of an unknown distribution density from observations", *Soviet Mathematics*, 3, 1559-1562

Chen, Y. and Tu, D.S. (1987). "Estimating the error rate in discriminant analysis by the delta, jackknife and bootstrap method", *Chinese Journal of Applied Probability and Statistics*, 3, 203-210

Chernick, M.R. (1985). "Application of bootstrap and other resampling techniques: evaluation of classifier performance", *Pattern Recognition Letters*, 3, 167-178

Chernick, M.R. (1986a). "Correction note to Application of bootstrap and other resampling techniques", *Pattern Recognition Letters*, 4, 133-142

Chernick, M.R. (1986b). "Estimation of error rate for linear discriminant functions by resampling: non-Gaussian populations", *Computing and Mathematics with Applications*, 15, 29-37

Chernoff, H. (1973). "Using faces to represent points in *k*-dimensional space graphically", *Journal of the American Statistical Association*, 68, 361-368

Choi, S.C. (1986). *Statistical methods of discrimination and classification - Advances in theory.* Pergamon Press

Christl, H.L. and Stock, S. (1973). "Ein dialogfähiges System zur Untersuchung simulierter Krankheitsbilder mittels verschiedener automatischer Klassifikations-verfahren", *J. d'Inform. Med.*, 1, 283-294

Clyma, J.A. and Lancaster, G. (1993). "Computer aided diagnosis of abdominal pain: implementation difficulties in the North Western Region", *Archives of Emergency Medicine*, 10, 314-320

Cochran, W.G. and Hopkins, C.E. (1961). "Some classification problems with multivariate qualitative data", *Biometrics*, 17, 10-32

Cochran, W.G. (1966). "Contribution to the discussion of the paper by M. Hills", *Journal of the Royal Statistical Society, Series B*, 28, 28-29

Cole T.J., Morley C.J., Thornton A.J., Fowler M.A. and Hewson, P.H. (1991). "A scoring system to quantify illness in babies under 6 months of age", *Journal of the Royal Statistical Society, series A*, 154(2), 287-304

Cornfield J. (1962). "Joint dependence of risk of coronary heart disease on serum cholestorol and systolic blood pressure: a discriminant function approach", *Fed. Americ. Socs. Exper. Biol. Proc. Suppl*, 11, 58-61

Cover T.M., and Hart P.E. (1967). "Nearest neighbour pattern classification", *IEEE Transactions on Information Theory*, IT-13, 21-27

Cover T.M., and Wagner T.J. (1976). "Topics in statistical pattern recognition", *Communications in cybernetics*, 10, 15-46

Cox, D.R. (1970). *The Analysis of Binary Data.* First Edition. London: Methuen

Cox, D.R. (1972). "The Analysis of Multivariate Binary Data", *Applied Statistics*, **21**, 113-120

Cox, D.R. and Snell, E.J. (1989). *Analysis of binary data*. Second Edition. London: Chapman and Hall

Cox, T.F. and Ferry, G. (1993). "Discriminant analysis using non-metric multidimensional scaling", *Pattern Recognition*, **26**(1), 145-153

Critchley, F. (1987). "Uncertainty in discrimination", Proc Conf DIANA II, Prague: *Mathematical Institute of the Czechoslovak Academy of Sciences*, 83-106

Critchley, F. and Vitiello, C. (1991). "The Influence of Observations on Misclassification Probability Estimates in Linear Discriminant-Analysis", *Biometrika*, **78**(3), 677-690

Crownover, R.M. (1991). "A least squares approach to linear discriminant analysis", *Siam Journal on Scientific and Statistical Computing*, **12**(3), 596-606

Curram, S.P. and Mingers, J. (1994). "Neural networks, decision tree induction and discriminant analysis - an empirical comparison", *Journal of the Operational Research Society*, **45**(4), 440-450

Cwik, J. (1989). "Estimating density ratio with application to discriminant analysis", *Communications in Statistical Theory and Methods*, **18**, 3057-3069

Dalgliesh, L.I. (1994). "Discriminant analysis - statistical inference using the jackknife and bootstrap procedures", *Psychological Bulletin*, **116**(3), 498-508

Dawber, T.R., Kannel, W.B. and Lyell, L.P. (1963). "An approach to longitudinal studies in a community; The Framingham study.", *Annals of the New York Academy of Sciences*, **107**, 539-556

Dawid, A.P. (1976). "Properties of diagnostic data distributions", *Biometrics*, **32**, 647-658

Devroye, L.P. (1982). "Any discrimination rule can have an arbitrarily bad probability of error for fi", ((?????)) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **4**, 154-157

Devroye, L.P. and Wagner, T.J. (1982). "Nearest neighbour methods in discrimination", In *Handbook of Statistics*, 2, Krishnaiah, P.R. and Kanal, L.N. (Eds.), Amsterdam: North Holland, 193-197

Devroye, L.P. (1987). *A course in density estimation*. Birkhauser

Devroye, L.P. (1988). "Automatic *Pattern Recognition*: a study of the probability of error", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **10**, 530-543

DiCiccio T.J. and Romano J.P. (1988). "A review of bootstrap intervals (with discussion).", *Journal of the Royal Statistical Association, Series B*, **50**, 338-370

Dickey J.M. (1968). "Smoothed estimates for multinomial cell probabilities", *Annals of Mathematical Statistics*, 39, 2, 561-6

Dillon, W.R. and Goldstein, M. (1978). "On the performance of some multinomial classification rules", *Journal of the American Statistical Association*, 73, 305-313

Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley

Dunsmore, I.R. (1966). "A Bayesian approach to classification", *Journal of the Royal Statistical Society. Series B* 28, 568-577

Efron, B. (1979). "Bootstrap methods: another look at the jacknife", *The Annals of Statistics*, 7, 1-26

Efron, B. (1982). "The jacknife, the bootstrap and other resampling plans", *SIAM* 3.

Efron, B. (1983). "Estimating the error rate of a prediction rule: improvement on cross-validation", *Journal of the American Statistical Association*, 78, 316-331

Efron, B. (1987). "Better bootstrap confidence intervals", (with discussion), *Journal of the American Statistical Association*, 82, 171-200

Efron, B. (1990). "More efficient bootstrap computations", *Journal of the American Statistical Association*, 85(409), 79-89

Elashoff, J.D., Elashoff, R.M. and Goldman, G.E. (1967). "On the choice of variables in classification problems with dichotomous variables", *Biometrika*, 54, 668-670

Epanechnikov, V.A. (1969). "Non-parametric estimation of a multivariate probability density", *Theor. Prob. Appl.*, 14, 153-158

Fahrmeir, L., Häußler, W., Tutz, G. (1984). *Multivariate Verfahren*.

Falbo, P. (1991). "Credit scoring by enlarged discriminant models", *Omega*, 19(4), 275-289

Feldmann, U., Schneider, B., Klinkers, H. and Haeckel, R. (1981). "A multivariate approach for the biometric comparison of analytical methods in clinical chemistry", *Journal of Clinical Chemistry and Clinical Biochemistry*, 19, 121-137

Feldmann, U. (1987). "Multivariater Methodenvergleich" <Multivariate method comparison>, *Lecture given at the 6th conference of the German Society for Clinical Chemistry on "Statistical problems in method comparisons"*, Berlin, 22.01.1987

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. 2nd edition, London: MIT Press

Fisher, R.A. (1936). "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, 7(2), 179-188

Fitzmaurice, G.M. and Hand, D.J. (1987). "A comparison of two average conditional error rate estimators", *Pattern Recognition Letters*, 6, 221-224

Fitzmaurice, G.M. (1990). "A Monte Carlo study of the 632 bootstrap estimator of error rate", *Journal of Classification*

Fix, E. and Hodges, J.L. (1951). "*Non-parametric discriminant analysis*", United States Air Force School of Aviation Medicine, Project Number 21-49-004, Reports. 4 & 11

Flick, T.E. (1990). "Pattern Classification using Projection Pursuit", *Pattern Recognition*, 23, 1367-1376

Frangos, C.C. (1984). "On jacknife, cross-validatory and classical methods of estimating a proportion", *Biometrika*, 71, 361-366

Freed, N. (1981). "A linear programming approach to the discriminant problem", *Decision sciences*, 12, 68-74

Friedman, J.H. (1984). "Projection pursuit - discussion", *The Annals of Statistics*, 13(2), 475-481

Friedman, J.H. (1987). "Exploratory projection pursuit", *Journal of the American Statistical Association*, 82, 249-266

Friedman, J.H. (1989). "Regularized discriminant analysis", *Journal of the American Statistical Association*, 84, 165-175

Fukunaga, K. (1971). "Estimation of classification error", *IEEE Transactions on Computing*, C20, 1521-1527

Fukunaga, K. and Hotstetler, L.D. (1973). "Optimization of k-nearest neighbour density estimates", *IEEE Transactions on Information Theory*, 19, 320-326

Fukunaga, K. and Kessel, D. (1973). "Nonparametric Bayes error estimation using unclassified samples", *IEEE Transactions on Information Theory*, 4, 434-440

Fukunaga, K. and Hayes, R.R. (1989a). "Effects of sample size in classifier design", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8), 873-885

Fukunaga, K. and Hayes, R.R. (1989b). "Estimation of classifier performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(10), 1087-1101

Fukunaga, K. and Hummels, D.M. (1989). "Leave-one-out procedures for nonparametric error estimates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(4), 421-423

Gallinari, P., Thiria, S. and Fogelmansoulie, F. (1988). "Multilayer perceptrons and data analysis", *IEEE*, 1, 391-399

Gallinari, P., Thiria, S., Badran, F. and Fogelmansoulie, F. (1991). "On the relations between discriminant analysis and multilayer perceptrons", *Neural Networks*, 4(3), 349-360

Ganesalingam, S. and McLachlan, G.J. (1980). "Error rate estimation on the basis of posterior probabilities", *Pattern Recognition*, 12, 405-413

Garson, G.D. (1991). "A comparison of neural network and expert systems algorithms with common multivariate procedures for analysis of social science data", *Social Science Computer Review*, 9(3), 399-434

Gessaman, M.P. (1972). "A comparison of some multivariate discrimination procedures", *Journal of the American Statistical Association*, 67, 468-472

Gilbert, E.S. (1968). "On discrimination analysis using qualitative variables", *Journal of the American Statistical Association*, 63, 1399-1412

Glick, N. (1972). "Sample based classification procedures derived from density estimators", *Journal of the American Statistical Association* 67(337) 116-122

Glick, N. (1973a). "Sample based multinomial classification", *Biometrics*, 29, 241-256

Glick, N. (1976). "Sample based classification procedures related to empiric distributions", *IEEE Transactions on Information Theory*, II-22(4), 454-461

Glick, N. (1978). "Additive estimators for probabilities of correct classification", *Pattern Recognition*, 10, 211-222

Gnanadesikan, R., Blashfield, R.K., Breiman, L., Dunn, O.J., Friedman, J.H., Fu, K.S., Hartigan, J.H., Kettenring, J.R., Lachenbruch, P.A., Olshen, R.A. and Rohlf, F.J. (1989). "Discriminant analysis and clustering: Panel on discriminant analysis, classification and clustering", *Statistical Science*, 4, 34-69

Goin, J.E. (1984). "Classification bias of the k-nearest neighbour algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 379-381

Goldstein, M. and Rabinowitz, M. (1975). "Selection of variables for the two-group multinomial classification problem", *Journal of the American Statistical Association*, 70, 776-781

Goldstein, M. and Dillon, W.R. (1977). "A stepwise discrete variable selection procedure", *Communications in Statistics, Theory and Materials*, 6, 1423-1436

Goldstein, M. and Wolf, E. (1977). "On the problem of bias in multinomial classification", *Biometrics*, 33, 325-31

Goldstein, M. and Dillon, W.R. (1978). *Discrete Discriminant Analysis*. New York: Wiley

Goldstein, M. and Dillon, W.R. (1984) *Multivariate Analysis - methods and applications*. New York: Wiley

Granville, V. and Rasson, J.P. (1995). "Multivariate discriminant analysis and maximum penalized likelihood density estimation", *Journal of the Royal Statistical Society*. B

Griffiths, P and Hill, I.D. (1985). *Applied Statistics Algorithms*. Chichester: Ellis Horwood Ltd. <published for the Royal Statistical Society>

Grizzle, J.E., Starmer, C.E. and Koch, G.G. (1969). "Analysis of categorical data by linear models", *Biometrics*, 25, 489-504

Grozinger, M., Freisleben, B. and Roschke, J (1994). "Comparison of a backpropagation network and a nonparametric discriminant analysis in the evaluation of sleep EEG data", *World Congress on Neural Networks, San Diego. International Neural Network Society Annual Meeting*, 1, 462-466

Gupta, Y.P., Bagci, P.K. and Rao, R.P. (1990). "A comparative analysis of the performance of alternative discriminant procedures: an application to bankruptcy prediction", *Journal of Information and Optimisation Sciences*, 11(3), 457-471

Habbema, J.D.F., Hermans, J. and van den Broek (1974). "A stepwise discriminant analysis programme using density estimation", In Compstat 1974, Proceedings of *Computational Statistics*, Vienna: Physica Verlag, 101-110

Hadzikadic, M. (1992). "Automatic design of diagnostic systems", *Artificial Intelligence in Medicine*, 4, 329-342

Haerting, J. (1983). "Special properties in selection performance of qualitative variables in discriminant analysis", *Biometrical Journal*, 25, 215-222

Hall, P. (1981a). "On nonparametric multivariate binary discrimination", *Biometrika*, 68(1), 287-294

Hall, P. (1981b). "Optimal near neighbour estimator for use in discriminant analysis", *Biometrika*, 68(2), 572-575

Hall, P. (1987). "On smoothing sparse multinomial data", *Australian Journal of Statistics*, 29, 19-37

Hall, P. (1988). "On nonparametric discrimination using density differences", *Biometrika*, 75, 541-547

Hall, P. (1995). *The Bootstrap and Edgeworth Expansion*. London: Springer Verlag

Hand, D.J, (1981). *Discrimination and classification*. New York: Wiley

Hand, D.J. (1982). *Kernel discriminant analysis*. New York: Research Studies Press.

Hand, D.J. (1983). "A comparison of two methods of discriminant analysis applied to binary data", *Biometrics*, 39, 683-694

Hand, D.J. (1986). "Recent advances in error rate estimation", *Pattern Recognition Letters*, 4, 335-346

Hand, D.J. (1992). "The diagnosis of disease", *Statistical Methods in Medical Research*

Hastie, T., Tibshirani, R. and Buja, A. (1994). "Flexible discriminant analysis by optimal scoring", *Journal of the American Statistical Association*, 89(428), 1255-1270

Hawkins, D.M. and Kass, G.V. (1982). "Automatic Interaction Detection", in Hawkins, D.M. (Ed), *Topics in Applied Multivariate Analysis*, 267-302, Cambridge Univ Press: Cambridge.

Hermans, J. (1981). "Use of posterior probabilities to evaluate methods of discriminant analysis", *Methods of Information in Medicine*, 20, 207-212

Hermans, J., Habbema, J.D.F. and Schäfer, J.R. (1982). "The ALLOC80 package for discriminant analysis", *Statistical Software Newsletter*, 8, 15-20

Hertz, J., Krogh, A. and Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation*. Redwood City: Addison-Wesley

Hillion, A., Masson, P. and Roux, C. (1988). "A non-parametric approach to linear feature extraction; application to classification of binary synthetic textures", *IEEE Comput. Soc. Press*, 2, 1036-1039

Hills, M. (1966). "Allocation rules and their error rates", *Biometrics*, 1, 1-31

Hills, M. (1967). "Discrimination and allocation with discrete data", *Journal of the Royal Statistical Society, Series C*, 16, 237-250

Hinkley, D.V. (1988). "Bootstrap methods", *Journal of the Royal Statistical Society, Series B*, 50(3), 321-337

Hinton, G.E. (1992). "How neural networks learn from experience", *Scientific American*, 267, 144-151

Hirst, D. (1988). "Applications of measures of uncertainty in discriminant analysis", In *Lecture Notes in Computer Science*, J. Kittler (Eds), Springer Verlag 487-496

Hjort, N.L. and Mohn, E. (1984). "Comparison of some contextual methods in remote sensing", *Environmental Research Institute of Michigan*, Ann Arbor, 1693-1702

Hjort, N.L. (1986). *Notes on the theory of statistical symbol recognition*. Report 778, Norwegian Computing Center, Oslo

Hogg, R.V. (1974). "Adaptive robust procedures: A partial review and some suggestions for future applications and theory", *Journal of the American Statisical Association*, 69(348), 909-923

Hora, S.C. and Wilcox, J.B. (1982). "Estimation of error rates in several-population discriminant analysis", *Journal of Marketing Research*, XIX, 57-61

Huberty, C.J. (1987). "Assessing predictive accuracy in discriminant analysis", *Multivariate Behavioural Research*, 22, 307-329

Hufnagl, P. (1985). "Fidelity estimation for a hierarchical classifier", *Biometrical Journal*, 6, 659-667

Jones, M.C. and Sibson, R. (1987). "What is projection pursuit?", *Journal of the Royal Statistical Society, Series A*, 150(1), 1-38

Kanal, L. (1974). "Patterns in *Pattern Recognition*: 1968-1974", *IEEE Transactions on Information Theory*, 20, 697-722

Kass, G.V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, 29, 119-127.

Kendall, M.G. (1971) *Rank corelation methods* (4th edition), Charles Griffin and Co., London and High Wycombe

Kharin, Y.S. (1990). "Robustness of discriminant analysis procedures - Survey", *Industrial Laboratory - USSR*, 56(10), 1236-1241

Kim, B.S. (1986). "A fast k nearest neighbour finding algorithm based on the ordered partition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8, 761-766

Kish, L. (1957). "Confidence intervals for clustered samples", *American Sociological Review*, 22, 154-165

Kittler, J. and Devijver, P.A. (1981). "An efficient estimator of Pattern Recognition system error probability", *Pattern Recognition*, 13, 245-249

Kittler, J. and Devijver, P.A. (1982). "Statistical properties of error estimators in performance assessment of recognition systems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4, 215-220

Konig, J. (1988). "Error rate estimation in discriminant analysis in the presence of variable selection", In *Expert Systems and Decision Support in Medicine. 33rd Annual Meeting of the GMDS EFMI Special Topic Meeting. Peter L. Reichertz Memorial Conference*, 162-168

Koffler, S.L. (1979). "Nonparametric discrimination procedures for non-normal distributions", *Journal of Statistical Computing and Simulation*, 8, 281-299

Krishnan, T. and Nandy, S.C. (1990). "Efficiency of discriminant analysis when initial samples are classified stochastically", *Pattern Recognition*, 23(5), 529-537

Kronmal, R.A. and Tarter, M. (1968). "The estimation of probability densities and cumulatives by Fourier series methods", *Journal of the American Statistical Association*, 63, 925-952

Krusinska, E. and Liebhart, J. (1988). "Robust selection of the most discriminative variables in the dichotomous problem with application to some respiratory disease data", *Biometrical Journal*, **30**, 295-303

Krusinska, E. and Liebhart, J. (1989). "Some further remarks on robust selection of variables in discriminant analysis", *Biometrical Journal*, 31, 227-233

Krzanowski, W.J. (1982). "Mixtures of continuous and categorical variables in discriminant analysis: a hypothesis testing approach", *Biometrics*, 38, 991-1002

Krzanowski, W.J., Jonathan, P., McCarthy, W.V. and Thomas, M.R. (1995). "Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data", *Applied Statistics*, 44, 101-115

Kuhnel, H. and Tavan, P. (1991). "A network for discriminant analysis", In *Artificial Neural Networks. Proceedings of the 1991 International Conference.* (Eds. Kohonen, T., Makisara, K., Simula, O. and Kangas, J.), ICANN-91, 2, 1053-1056

Kullback, S. and Leibler, A. (1951). "On information and sufficiency", *Annals of Mathematical Statistics*, 22, 79-86

Kurita, T., Asch, H. and Otsu, N. (1994). "Nonlinear discriminant features constructed by using outputs of multilayer perceptron", In *ISSIPNN '94. 1994 International Symposium on Speech, Image Processing and Neural Networks Proceedings.* (Cat. No.94TH0638-7), 2, 417-420

Kuss, E., Tryba, M., Kürzl, R. and Ulsenheimer, K. (1991). "Welcher Nutzen und welcher Schaden kann von Screening- und Routineuntersuchungen erwartet werden und von deren Unterlassung?", <Benefit and harm to be expected from screening and routine tests or their omission>, *Zeitschrift für Geburtshilfe und Frauenheilkunde*, **51**, (415-430)

Lachenbruch, P.A. (1967). "An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis", *Biometrics*, 23, 639-645

Lachenbruch, P.A. (1968). "Estimation of error rates in discriminant analysis", *Technometrics*, 10, 1-11

Lachenbruch, P.A. (1975). "Discriminant analysis". New York: Hafner Press

Lack, H.N. (1986). "Loglinear analysis of stillbirth incidence and social risk factors", *Paper presented at 31st German stat. soc. meeting in Göttingen*

Lack, H.N. (1987). "Distance measures in discrete discriminant analysis", *Poster presented at 32nd German Stat. Soc. Meeting in Tübingen*

Lack, H.N. (1987). *Prediction of stillbirths: a comparative discriminant approach.* M.Sc. thesis, Sheffield City Polytechnic

Lack, H.N. (1988). "Non-parametric methods for discriminant analysis of discrete data", *lecture given at Dept. of Biometrics Hanover medical school*

Lack, H.N. (1992). "Qualitätsindikatoren - Entwicklung einer geeigneten Darstellung", <Indicators of quality - development of a suitable form of presentation>, in: *BPE Jahresbericht 1991*, Kommission für Perinatologie und Neonatologie der Bayerischen Landesärztekammer und der Kassenärztlichen Vereinigung Bayerns <annual report of Bavarian Perinatal Survey Steering Committee>

Lack, H.N. (1993). "Schwangerenvorsorge", <Antenatal care>, In *BPE Jahresbericht 1993*, Kommission für Perinatologie und Neonatologie der Bayerischen Landesärztekammer und der Kassenärztlichen Vereinigung Bayerns <annual report of Bavarian Perinatal Survey Steering Committee>

Lack, H.N. (1994). "Läβt sich eine drohende Frühgeburt vorhersagen ?", <Are imminent premature deliveries predictable ?>, In *BPE Jahresbericht 1993*, Kommission für Perinatologie und Neonatologie der Bayerischen Landesärztekammer und der Kassenärztlichen Vereinigung Bayerns <annual report of Bavarian Perinatal Survey Steering Committee>

Lancaster, H.O. (1969). "Contingency tables of higher dimensions", In *The Chisquared Distribution*, Lancaster, H.O., New York: Wiley 253-281

Lark, R.M. (1994). "Sample size and class variability in the choice of a method of discriminant analysis", *International Journal of Remote Sensing*, 15(7), 1551-1555

Lawoko, C.R.O. and McLachlan, G.J. (1989). "Bias associated with the discriminant analysis approach to the estimation of mixing proportions", *Pattern Recognition* 22 (6), 763-766

Lazarsfeld, P.F. (1956). *Some observations on dichotomous systems. Duplicated Report.* New York: Department of Sociology, Columbia University.

Lazarsfeld, P.F. (1960). "Latent Structure Analysis", In *Psychology, the State of a Science.* Vol 3, New York: John Wiley

Lazarsfeld, P.F. (1961). "The algebra of dichotomous systems", In *Studies in Item Analysis and Prediction*, H. Solomon (Ed.). Stanford: Stanford University Press, 111-157

Legitimus, D. and Schwab, L. (1991). "Experimental comparison between neural networks and classical techniques of classification applied to natural underwater transients identification", *IEEE Conference on Neural Networks for Ocean Engineering* (Cat. No.91CH3064-3), 113-120

Lesaffre, E. (1989). "Estimation of error rate in multiple group logistic discrimination. The approximate leaving-one-out method", *Communications in Statistical Theory and Methods*, 18, 2989-3007

Lesaffre, E. (1989). "Multiple-group logistic regression diagnostics", *Journal of the Royal Statistical Society Series C*, 38, 425-440

Lindgren, B.W. (1976). *Statistical Theory*. 3rd edition, New York: Macmillan

Lissack, T. and Fu, K.S. (1976). "Error estimation in *Pattern Recognition* via L^a-distance between posterior density functions. *IEEE Transactions on Information Theory*, 22, 34-35

Little, R.J.A. (1978). "Consistent regression methods for discriminant analysis with incomplete data", *Journal of the American Statistical Association*, 73, 319-322

Loftsgaarden, D.O. and Quesenberry, C.P. (1965). "A nonparametric estimate of a multivariate density function", *Annals of Mathematical Statistics*, 36, 1049-1051.

Loh, W.Y. and Vanichsetakul, N. (1988). "Tree-structured classification via generalized discriminant-analysis", *Journal of the American Statistical Association*, 83(403), 715-725

Loh, W.Y. (1988). *Fast Algorithm for Classification Trees (FACT)*. Users guide to version 1.1, Department of Statistics, University of Wisconsin, Madison

Loh, W.L. (1995). "On linear discriminant analysis with adaptive ridge classification rules", *Journal of Multivariate Analysis*, 53(2), 264-278

Lowe, D. and Webb, A.R. (1991). "Optimized feature extraction and the Bayes decision in feed-forward classifier networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4), 355-364

McCullagh P. and Nelder J.A. (1989). *Generalised linear models*. 2nd edition, London: Chapman and Hall.

McKenzie, D.P., McGorry, P.D., Wallace, C.S., Low, L.H., Copolow, D.L. and Singh, B.S. (1993). "Constructing a minimal diagnostic decision tree", *Methods of Information in Medicine*, 32(2), 161-166

McLachlan, G.J. (1976). "A criterion for selecting variables for the linear discriminant function", *Biometrics*, 32, 529-515

McLachlan, J. (1977). "The bias of sample based posterior probabilities", *Biometrical Journal*, 19, 421-426

McLachlan, G.J. (1980). "The efficiency of Efron's bootstrap approach to error rate estimation in discriminant analysis", *Journal of Statistical Computing and Simulation*, 11, 273-279

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley

Marcotte, P., Marquis, G. and Savard, G. (1995). "A new implicit enumeration scheme for the discriminant analysis problem", *Computers and Operations Research*, 22(6), 625-639

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press

Marron, J.S. (1989). "Comments on a data based bandwidth selector", *Computational Statistics and Data Analysis*, 8, 155-170

Marshall, R.J. (1986). "Partitioning methods for classification and decision making in medicine", *Statistics in Medicine*, 5, 517-526

Martin, D.C. and Bradley, R.A. (1972). "Probability models, estimation and classification for multivariate dichotomous populations", *Biometrics*, 28, 203-221

Matusita, K. (1955). "Decision rules, based on the distance, for problems of fit, two samples, and estimation", *Annals of the Institute of Mathematical Statistics*, 26, 631-640

Matusita, K. (1956). "Decision rule, based on distance, for the classification problem", *Annals of Mathematical Statistics*, 8, 67-77

Meulepas, E. (1990). "On a criterion for omitting variables in discriminant analysis", *Biometrics*, 46(4), 1181-1183

Minsky and Papert (1969). *Perceptrons*. Cambridge Massachusets: MIT-Press

Mitra, S. and Kuncheva, L.I. (1995). "Improving classification performance using fuzzy MLP and two-level selective partitioning of the feature space", *Fuzzy Sets and Systems*, 70(1), 1-13

Molnar, L.D. (1991). "Loglindi - log-linear discriminant analysis", *American Statistician*, 45(4), 339-339

Moore, D.H. II (1973). "Evaluation of five discrimination procedures for binary variables", *Journal of the American Statistical Association*, 68(342), 399-404

Moore, D.S., Whitsell, S.J. and Lundgrebe, D.A. (1976). "Variance comparisons for unbiased estimators of probability of correct classification", *IEEE Transactions on Information Theory*, 22, 102-105

Moore, M. (1982). *Discriminant analysis with discrete data*. BSc., Part 3 Project, Sheffield City Polytechnic

Morgan, J.N. and Messenger, R. (1973). *THAID: a sequential search program for the analysis of nominal scale dependent variables*. Ann Arbor: Institute for Social Research

Müller, A. and Neumann, J. (1991). "Classification with neural networks", In *Classification, Data Analysis, and Knowledge Organization - Models and Methods with Applications* (Eds. Bock, H.H. and Ihm, P.). Proceedings of the 14th Annual Conference of the Gesellschaft für Klassifikation e.V, 32-42

Murray, G.D. and Titterington, D.M. (1978). "Estimation problems with data from a mixture", *Applied Statistics*, 27, 325-334

Myers, R.H. (1986). *Classical and Modern Regression with Applications*. Boston: Duxbury Press

Myles, J.P. (1990). "The multi-class metric problem in nearest neighbour discrimination rules", *Pattern Recognition*, 23, 1291-1291

Nath, R. and Jones, T.W. (1988). "A variable selection criterion in the linear programming approaches to discriminant analysis", *Decision Sciences*, 19(3), 554-563

Nelder, J.A. and Wedderburn, R.W.M. (1972). "Generalised linear models", *Journal of the Royal Statistical Society, Series A*, 135, 370-384

Niemann, H. (1988). "An efficient branch and bound nearest neighbour classifier", *Pattern Recognition Letters*, 7, 67-72

Odom, M.D., Sharda, R. (1990). "A neural network model for bankruptcy prediction", *1990 International Joint Conference on Neural Networks - IJCNN 90*, Part 2 (of 3), Jun 17-21, NJ, USA (IEEE cat n 90CH2879-5), 163-168

Ogorman, T.W. and Woolson, R.F. (1991). "Variable selection to discriminate between 2 groups - stepwise logistic regression or stepwise discriminant analysis", *American Statistician*, 45(3), 187-193

Ohmann, C., Platen, C., Belenky, G., Franke, C., Otterbeck, R., Lang, K. and Röhrer, H.D. (1995). "Expertensystem zur Unterstützung von Diagnosestellung und Therapiewahl bei akuten Bauchschmerzen", <An expert system for assisting diagnosis and choice of therapy in the case of acute abdominal pain>, *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, 26(3), 262-274

Olkin, I. and Spiegelman, C.H. (1987). "A semiparametric approach to density-estimation", *Journal of the American Statistical Association*, 82(399), 858-865

Oppel, U.G. (1990), "HUGIN, A software package for the construction of expert systems based on causal probabilistic networks", *Proceedings of the workshop "Uncertainty in knowledge based systems"*, FAW in Ulm, July 8-13, 250-260

Osman, H. and Fahmy, M.M. (1994). "On the discriminatory power of adaptive feed-forward layered networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8), 837-842

Ott, J. and Kronmal, R.A. (1976). "Some classification procedures for multivariate binary data using orthogonal functions", *Journal of the American Statistical Association*, 71, 391-399

Parzen, E. (1962). "On estimation of a probability density function and mode", *Annals of Mathematical Statistics*, 33, 1065-1076.

Patuwo, E., Hu, M.Y. and Hung, M.S. (1993). "Two group classification using neural networks", *Decision Sciences*, 24(4), 825-845

Pfeiffer, K.P. (1987). "Which discriminant function should be used?", EFMI - European Federation for Medical Informatics, *Proceedings of the Seventh International Congress*, 3, 1257-1261

Pipberger, H.V., Klingeman, J.D. and Cosma, J. (1968). "Computer evaluation of statistical properties of clinical information in the differential diagnosis of chest pain", *Methods of Information in Medicine*, 7(2), 79-92

Plackett, (1974). in Fienberg, S: *The analysis of cross classified categorical data*. MIT Press

Preisendorfer, R.W., Mobley, C.D. and Barnett, T.P. (1988). "The principal discriminant method of prediction: theory and evaluation", *Journal of Geophysical Research*, 93, 10815-10830

Pridmore, M. (1985). *Discrimination with discrete data*. B.Sc., Part 3 Project, Sheffield City Polytechnic

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufman: San Mateo, CA.

Rao, C.R. (1960), "Multivariate analysis: An indispensable statistical aid in applied research", *Sankhya*, 22, 317-338

Rao, C.R. (1962), "Use of discriminant and allied functions in multivariate analysis", *Sankhya*, A22, 149-154

Raveh, A. (1989). "A nonmetric approach to linear discriminant analysis", *Journal of the American Statistical Association*, 85(405), 176-183

Rehm, J. (1990). "Zur Schätzung von linearen Modellen mit fehlenden Werten in der Epidemiologie" <On estimation of linear models in the presence of missing data in epidemiology>, *personal communication*

Reiss, I.L., Banwart, A. and Foreman, H. (1975). "Premarital contraceptive usage: a study and some theoretical explanations", *Journal of Marraige and the Family*, 37, 619-630

Reibnegger, G., Weiss, G., Wernerfelmayer, G., Judmaier, G. and Wachter, H. (1991). "Neural networks as a tool for utilising laboratory information - comparison with linear discriminant analysis and with classification and regression trees", *Proceedings of the National Academy of Sciences of the United States of America*, 88(24), 1426-1430

Ripley, B.D. (1993). "Statistical aspects of neural networks", In *Networks and Chaos - Statistical and Probabilistic Aspects*. (Eds. Barndorff-Nielsen, O.E., Jensen, J.L. and Kendall, W.S.), London: Chapmann and Hall, 40-123

Ripley, B.D. (1994). "Neural networks and related methods for classification", *Journal of the Royal Statistical Society, Series B*, 56(3), 409-456

Ritter, H., Martinez, T. and Schulten, K. (1991). *Neuronale Netze - Eine Einführung in die Neuroinformatik selbstorganisierender Netzwerke*. Bonn: Addison-Wesley

Rosenberg, M. (1962). "Test factor standardisation as a method of interpretation.", *Social Forces*, 41, 53-61

Rosenblatt, F. (1958). "The Perceptron: a Probabilitstic Model for Information Storage and Organization in the Brain", *Psychological Review*, 65, 386-408.

Rosenblatt, M. (1956). "Remarks on some nonparametric estimates of a density function", *Annals of Mathematical Statistics*, 27, 832-837.

Rosenblatt, M. (1971). "Curve estimates", *Annals of Mathematical Statistics*, 42, 1815-1842

Rubin, D.B. (1976). "Estimation of missing data", *Biometrika*, 63, 581-589

Rubin, P.A. (1990). "A comparison of linear programming and parametric approaches to the two-group discriminant problem", *Decision Sciences*, 21(2), 373-386

Ruizvelasco, S. (1991). "Asymptotic efficiency of logistic regression relative to linear discriminant analysis", *Biometrika*, 78(2), 235-243

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986). "Learning representations by back-propagating errors", *Nature*, 323, 533-536

Sanchez, M.S. and Sarabia, L.A. (1995). "Efficiency of Multilayered Feedforward Neural Networks on Classification in Relation to Linear Discriminant Analysis, Quadratic Discriminant Analysis and Regularized Discriminant Analysis", *Chemometrics and Intelligent Laboratory Systems*, 28(2), 287-303

Sang Bin Lee and Seung Hyun Oh (1990). "A comparative study of recursive partitioning algorithm and analog concept learning system", *Expert Systems with Applications*, 1(4), 403-416

Sapra, S.K. (1991). "A connection between the logit model, normal discriminant analysis, and multivariate normal mixtures", *American Statistician*, 45(4), 265-268

Sarle, W.S. (1994). "Neural networks and statistical methods", *Proceedings of the 19th SAS users group international conference*, Texas

SAS Institute Inc. (1986). *SAS Procedures guide Version 6.* Third Edition, SAS Institute Inc., Cary, NC

Schervish, M.J. (1981). "Asymtotic expansion for correct classification rates in discriminant analysis", *Annals of Statistics*, 9, 1002-1009

Schewe, S., Herzer, P. and Krüger, K. (1990). "Prospective application of an expert system for the medical history of joint pain", *Klinische Wochenschrift*, 68, 466-471

Schoener, T.W. (1968). "The *anolis* lizards of Bimini: resource partitioning in a complex fauna", *Ecology*, 49, 704-726

Schwartz, S.C. (1967). "Estimation of probability density by orthogonal series", *Annals of Mathematical Statistics*, 38, 1261-1265

Sewell, W.H. and Shah, V.P. (1968). "Social class, parental encouragement and educational aspirations", *American Journal of Sociology*, 73, 559-572

Shao, J. and Tu, D. (1995). *The Jacknife and Bootstrap.* London: Springer Verlag

Shih, Y.S. (1994). *Partially Adaptive Classification Trees (PACT).* Users guide to version 1.1, Department of Mathematics, National Chung Cheng University of Taiwan

Shino, P.H. and Klebanoff, M.A. (1993): "Meta analysis of predictors of premature delivery", *Clinics in Perinatology*, 20(1), 107-125

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* London: Chapman and Hall

Silverman, B.W. and Jones, M.C. (1989). "Fix, E. and Hodges, J.L. (1951): An important contribution to nonparametric discriminant analysis and density estimation", Commentary on Fix and Hodges (1951), *International Statistical Review*, 57(3), 233-247

Siotani, M. (1982). "Large sample approximations and asymptotic expansions of classification statistics", In *Handbook of Statistics.* Volume 2, Krishnaiah, P.R. and Kanal, L.N., 61-100

Smith, C.A.B. (1947). "Some examples of discriminant analysis", *Annals of Eugenics*, 18, 272-283

Snappin, S.M. and Knoke, J.D. (1989). "Estimation of error rates in discriminant analysis with selection of variables", *Biometrics*, 45(1), 289-299

Snorrason, O. and Garber, F.D. (1992). "Evaluation of non-parametric discriminant analysis techniques for radar signal feature selection and extraction", *Optical Engineering*, 31(12), 2608-2617

Solomon, H. (Ed.) (1961). *Studies in item analysis and prediction*. Stanford University Press

Solow, A.R. (1990). "A randomization test for misclassification probability in discriminant analysis", *Ecology*, 71(6), 2379-2382

Sonquist, J.A. and Morgan J.N. (1964). *The detection of interaction effects*. Ann Arbour: Institute for Social Research University of Michigan

Specht, D.F. (1967). "Generation of polynomial discriminant functions for *Pattern Recognition*", *IEEE Transactions on Electronics and Computing*, EC-16, 308-319

SPSS, Inc. (1986). *SPSS-X User's Guide*. Second Edition, Chicago: Author

Stam, A. and Joachimsthaler, E.A. (1990). "A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem", *European Journal of Operational Research*, 46(1), 113-122

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society Series B*, 36, 111-147

Szczesny, W. (1991). "On the performance of a discriminant function", *Journal of Classification*, 8, 201-215

Tanaka, H., Ishibuchi, H. and Yoshikaw, S. (1994). "Discriminant analysis based on exponential possibility distributions", *IEEE*, 2, 802-807

Tapia, R.A. and Thompson, J.R. (1978). *Nonparametric Probability Density Estimation*. John Hopkins Series in the Mathematical Sciences No. 1, Baltimore, Maryland: John Hopkins University Press

Tarter, M. E. and Kronmal, R. A. (1970). "On multivariate density estimates based on orthogonal expansions", *Annals of Mathematical Statistics*, 4, 718-722

Tarter, M. E. and Kronmal, R.A. (1976). "An introduction to the implementation and theory of nonparametric density estimation", *American Statistician*, 30, 105-112

Thorndike, R.L. and Hagen, E.P. (1959). *Ten Thousand Careers*. New York: John Wiley.

Titterington, D.M. (1977). "Analysis of incomplete multivariate binary data by the kernel method", *Biometrika*, 64, 455-460

Titterington, D.M. (1980). "A comparative study of kernel-based density estimates for categorical data", *Technometrics*, 22(2), 259-68

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skeene, A.M., Habbema, J.D.F. and Gelpke, G.J. (1981). "Comparison of discrimination techniques applied to a complex dataset of head injured patients", *Journal of the Royal Statistical Society, Series A*, 144(2), 145-175

Todeschini, R. and Marengo, E. (1992). "Liner discriminant classification tree: a user driven multicriteria method", *Chemometrics and Intelligent Laboratory Systems*, 16(1), 25-35

Toussaint, G. T. (1974). "Bibliography on estimation of misclassification", *IEEE Transactions on Information Theory*, 20, 472-479

Trampisch, H.J. (1976). "A discriminant analysis for qualitative data with interactions", *Computer Programs in Biomedicine*, **6**, 50-60

Trampisch, H.J. (1983). "On the performance of some classification rules for qualitative data for simulated underlying distributions", *Biometrical Journal*, 25(7), 689-698

Tutz, G. (1988). "Smoothing for discrete kernels in discrimination", *Biometrical Journal*, 30, 729-739

Tutz, G. (1989). "On cross-validation for discrete kernel estimates in discrimination", *Communications in Statistical Theory and Methods*, 18(11), 4145-4162

Valova, D., Drska, Z., Polankova, M. and Malkova, A. (1993). "Comparison of discriminant analysis and probabilistic expert system in VCG data classification", *Physiological Research*, 42(2), 91-93

Victor, N., Trampisch, H.J. and Zentgraf, R. (1974). "Diagnostic rules for qualitative variables with interactions", *Methods of Information in Medicine*, **13**, 184-186

Victor, N. (1976). "Probleme der Auswahl geeigneter Zuordnungsregeln bei unvollständiger Information, insbesondere für kategoriale Daten", *Biometrics*, 32, 571-585

Villarroya, A., Rios, M. and Oller, J.M. (1995). "Discriminant analysis algorithm based on a distance function and on a Bayesian decision", *Biometrics*, 51, 908-919

Vlachonikolis, I.G. (1990). "Predictive discrimination and classificiation with mixed binary and continuous variables", *Biometrika*, 77, 657-662

Wakeley, P.C. (1954). "Planting the southern pines", *U.S. Dept. Agr. Forest. Serv. Agr. Monogr.*, 18, 1-233

Wang, M.C. and van Ryzin, J. (1981). "A class of smooth estimators for discrete distributions", *Biometrika*, 68(1), 301-309

Webb, A.R. and Lowe, D. (1990). "The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis", *Neural Networks*, 3(4), 367-375

Wegman, E.J. (1972a). "Non-parametric probability density estimation. I", *Technometrics*, 14, 533-546

Wegman, E.J. (1972b). "Non-parametric probability density estimation. II", *Journal of Statistical Computing and Simulation*, 1, 225-246

Weisberg, S. (1985). *Applied Linear Regression*. New York: Wiley

Weiss S.M. and Kulikowski C.A. (1991). *Computer systems that learn*, San Mateo, C.A.: Morgan Kaufmann

Werbos, P. (1974). *Beyond Regression: New Tools for Prediaction and Analysis in the Behavioual Sciences*. Ph.D. thesis, Harvard University Committe on Applied Mathematics

Wernecke, K.D. (1986). "The use of combined classifiers in medical functional diagnosis", *Biometrical Journal*, 28, 81-88

Wernecke, K.D., Haerting, J., Kalb, G. and Stürzebecher, E. (1989). "On model choice in discrimination with categorical variables", *Biometrical Journal*, 31(3), 289-296

Wernecke, K.D. (1992). "A coupling procedure for the discrimination with mixed data", *Biometrics*, 48, 497-506

Whittle, P. (1958). "On the smoothing of probability density functions", *Journal of the Royal Statistical Society, Series B*, 20, 334-343

Williams, B.K., Titus, K. and Hines, J.E. (1990). "Stability and bias of classification rates in biological applications of discriminant analysis", *Journal of Wildlife Management*, 54(2), 331-341

Wong, P.M., Jian, F.X. and Taggart, I.J. (1995). "A Critical Comparison of Neural Networks and Discriminant Analysis in Lithofacies, Porosity and Permeability Predictions", *Journal of Petroleum Geology*, 18(2), 191-206

Wyman, F.J., Young, D.M. and Turner, D.W. (1990). "A comparison of asymptotic error rate expansions for the sample linear discriminant function", *Pattern Recognition*, 23(7), 775-783

Yoon, Y.O., Swales, G. and Margavio, T.M. (1993). "A comparison of discriminant analysis versus artificial neural networks", *Journal of the Operational Research Society*, 44(1), 51-60

Zentgraf, R. (1975). "A note on Lancasters definition of higher order interactions", *Biometrika*, 62(2), 375-378

Appendix - Detailed list of results

The "results" chapters 13 and 14 contain only summary tables. All other tables with detailed results appear in this appendix. Chapters 13 and 14 may be read without reference to the appendix yet for a comprehensive view the following tables should be consulted. All tables generally relate to one of the three criteria $\varepsilon_{counting}$, $\varepsilon_{posterior}$ and η and show dataset down the vertical dimension and procedure across the horizontal. Each cell then relates to a single statistic for the particular data by procedure combination. Where appropriate suitable averages of these statistics have been calculated. Tables presenting such averaged statistics generally show all three criteria jointly for comparison. Statistics are averaged over data set, procedures or both.

Estimates of hold-out based expected values of performance criteria are listed by data set and discriminant procedure in appendix A. Data sets are divided into *real* and *artificial* ones, discriminant procedures are divided into *direct* and *indirect* procedures throughout appendix E. Appendix B tabulates variability of performance criteria averaged over data sets and over discriminant procedure. The expected standard errors for all three performance criteria appear in appendix C in three groups: for each data set - procedure combination, averaged over data sets or procedures and averaged over both. Appendices D and E give estimates of expected bias for the *conditional* and *unconditional* performance criteria respectively in the same grouping as for appendix C. Variations of performance criteria with respect to degree of *discreteness* of a given data set are given in appendix F. Detailed results for performance criteria in relation to varying classification thresholds τ are shown in appendix G.

Appendix A - Expectation of performance criteria

expectation of hold-out based estimates of err(counting)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dh1
type	pred	data											
real	dichot.	BREAST	0.340	0.341	0.340	0.343	0.341	0.354	0.344	0.353	0.344	0.343	0.360
		CESAR4	0.123	0.124	0.123	0.135	0.134	0.146	0.127	0.136	0.140	0.139	0.140
		GRADE	0.240	0.245	0.240	0.253	0.253	0.853	0.293	0.520	0.360	0.567	0.597
		LIZARD	0.111	0.113	0.114	0.126	0.111	0.140	0.113	0.141	0.109	-	0.169
		VIRGIN	0.200	0.207	0.209	0.204	0.209	0.220	0.210	0.205	0.204	0.297	0.204
	polytom.	CHD	-	-	-	0.069	0.069	-	0.069	0.340	0.270	0.322	0.510
		COLLEGE	-	-	-	0.201	0.200	-	0.203	0.274	0.215	0.215	0.234
		CREDIT	-	-	-	0.200	0.252	0.257	0.292	0.302	0.206	0.200	0.410
		EDUC	-	-	-	0.311	0.311	-	0.311	0.713	0.617	0.610	0.791
		ESTEEM	-	-	-	0.303	0.303	0.303	0.303	0.401	0.432	0.427	0.570
		IRIS	-	-	-	0.144	0.120	0.377	0.170	0.160	0.234	0.351	0.170
		KRETSCHM	-	-	-	0.360	0.309	0.355	0.424	0.366	0.204	0.205	0.414
		VOTING	-	-	-	0.167	0.172	-	0.169	0.104	0.160	0.309	0.103
artif.	dichot.	DILLON	0.004	0.005	0.005	0.006	0.090	0.200	0.091	0.496	0.094	0.420	0.366
		MA435300	0.440	0.443	0.443	0.370	0.420	0.504	0.369	0.451	0.301	-	0.366
		MA435301	0.435	0.441	0.441	0.360	0.416	0.540	0.346	0.439	0.293	-	0.345
		MA435302	0.399	0.400	0.403	0.303	0.407	0.472	0.374	0.399	0.313	-	0.373
		MA435303	0.402	0.400	0.399	0.370	0.377	0.460	0.356	0.400	0.302	-	0.363
		MA435304	0.390	0.405	0.401	0.346	0.366	0.443	0.339	0.405	0.200	-	0.341
		MA435305	0.432	0.427	0.420	0.357	0.419	0.497	0.350	0.427	0.202	-	0.346
		MA435306	0.400	0.403	0.401	0.399	0.401	0.450	0.377	0.401	0.290	-	0.375
		MA435307	0.433	0.425	0.432	0.359	0.393	0.495	0.352	0.430	0.294	-	0.349

(CONTINUED)

Table A-1: Counting based error rates

expectation of hold-out based estimates of err(counting)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435388	0.472	0.468	0.472	0.326	0.448	0.529	0.325	0.478	0.276	-	0.327
		MA435389	0.372	0.372	0.369	0.348	0.374	0.462	0.341	0.369	0.277	-	0.336
	polytom.	BANANA	-	-	-	0.037	0.191	0.215	0.036	0.183	-	0.163	0.037
		INTERAC1	-	-	-	0.069	0.378	0.408	0.060	0.378	0.068	0.060	0.068
		NORMAL01	-	-	-	0.284	0.231	-	0.374	0.229	0.168	0.164	0.386
		NORMAL02	-	-	-	0.394	0.386	-	0.398	0.447	0.244	0.249	0.402
		NORMAL03	-	-	-	0.020	0.017	-	0.032	0.063	0.016	0.016	0.032
		NORMAL11	-	-	-	0.039	0.037	0.083	0.214	0.056	0.012	0.009	0.289
		NORMAL12	-	-	-	0.040	0.032	0.067	0.120	0.070	0.015	0.015	0.126
		NORMAL13	-	-	-	0.021	0.018	0.043	0.065	0.055	0.015	0.015	0.065
		NORMAL14	-	-	-	0.041	0.034	0.077	0.050	0.053	0.024	0.025	0.047
		NORMAL15	-	-	-	0.052	0.051	0.055	0.061	0.059	0.039	0.042	0.061
		NORMAL16	-	-	-	0.061	0.044	0.059	0.054	0.056	0.044	0.044	0.057
		NORMAL17	-	-	-	0.056	0.052	0.198	0.058	0.167	0.055	0.055	0.057
		POISSON	-	-	-	0.075	0.069	-	0.079	0.078	0.064	0.064	0.080

Table A-1: Counting based error rates

expectation of hold-out based estimates of err(posterior_1)			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lgl	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.283	0.282	0.282	0.348	0.322	0.322	0.332	0.331	0.338	0.338	0.333
		CESAR4	0.119	0.120	0.120	0.110	0.072	0.315	0.110	0.124	0.115	0.119	0.126
		GRADE	0.215	0.215	0.213	0.220	0.202	0.207	0.192	0.186	0.219	0.155	0.239
		LIZARD	0.110	0.109	0.106	0.127	0.103	0.110	0.109	0.119	0.094	-	0.106
		VIRGIN	0.198	0.200	0.196	0.223	0.144	0.238	0.201	0.222	0.188	0.218	0.236
	polytom.	CHD	-	-	-	0.069	0.069	-	0.069	0.060	0.071	0.069	0.060
		COLLEGE	-	-	-	0.207	0.106	-	0.197	0.198	0.199	0.200	0.192
		CREDIT	-	-	-	0.195	0.230	0.360	0.170	0.174	0.167	0.177	0.175
		EDUC	-	-	-	0.311	0.314	-	0.312	0.313	0.311	0.310	0.466
		ESTEEM	-	-	-	0.304	0.303	0.270	0.303	0.304	0.304	0.301	0.301
		IRIS	-	-	-	0.113	0.052	0.379	0.132	0.125	0.083	0.100	0.096
		KRETSCHM	-	-	-	0.230	0.194	0.411	0.275	0.134	0.163	0.193	0.242
		VOTING	-	-	-	0.100	0.150	-	0.169	0.171	0.175	0.179	0.177
artif.	dichot.	DILLON	0.091	0.089	0.092	0.082	0.087	0.153	0.076	0.073	0.084	0.093	0.081
		MA435300	0.327	0.327	0.321	0.332	0.371	0.365	0.200	0.304	0.260	-	0.266
		MA435301	0.267	0.256	0.250	0.309	0.359	0.376	0.265	0.225	0.283	-	0.240
		MA435302	0.222	0.210	0.210	0.320	0.336	0.307	0.200	0.206	0.296	-	0.290
		MA435303	0.250	0.251	0.252	0.306	0.321	0.309	0.200	0.220	0.274	-	0.200
		MA435304	0.266	0.266	0.266	0.310	0.331	0.370	0.270	0.276	0.277	-	0.254
		MA435305	0.269	0.270	0.270	0.313	0.349	0.376	0.255	0.250	0.239	-	0.200
		MA435306	0.232	0.233	0.236	0.315	0.336	0.406	0.200	0.265	0.282	-	0.265
		MA435307	0.202	0.294	0.205	0.316	0.357	0.365	0.263	0.201	0.269	-	0.249

(CONTINUED)

Table A-2: Posterior based error rates

expectation of hold-out based estimates of err(posterior_1)			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435308	0.313	0.316	0.319	0.389	0.368	0.368	0.246	0.239	0.239	-	0.277
		MA435309	0.186	0.187	0.188	0.285	0.298	0.386	0.238	0.262	0.234	-	0.228
	polytom.	BANANA	-	-	-	0.046	0.169	0.185	0.037	0.048	-	0.033	0.039
		INTERAC1	-	-	-	0.073	0.389	0.282	0.060	0.066	0.069	0.069	0.060
		NORMAL01	-	-	-	0.169	0.214	-	0.239	0.212	0.168	0.144	0.168
		NORMAL02	-	-	-	0.328	0.277	-	0.258	0.278	0.223	0.223	0.218
		NORMAL03	-	-	-	0.035	0.016	-	0.026	0.018	0.028	0.022	0.021
		NORMAL11	-	-	-	0.261	0.024	0.481	0.194	0.079	0.067	0.056	0.098
		NORMAL12	-	-	-	0.288	0.017	0.366	0.184	0.044	0.031	0.049	0.063
		NORMAL13	-	-	-	0.099	0.022	0.339	0.051	0.033	0.011	0.015	0.013
		NORMAL14	-	-	-	0.061	0.034	0.275	0.038	0.033	0.032	0.011	0.038
		NORMAL15	-	-	-	0.057	0.036	0.234	0.049	0.038	0.065	0.028	0.046
		NORMAL16	-	-	-	0.062	0.058	0.195	0.053	0.042	0.041	0.053	0.078
		NORMAL17	-	-	-	0.057	0.003	0.072	0.055	0.041	0.046	0.065	0.054
		POISSON	-	-	-	0.068	0.052	-	0.062	0.065	0.066	0.054	0.071

Table A-2: Posterior based error rates

expectation of hold-out based estimates of eta			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lgi	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.689	0.689	0.689	0.659	0.668	0.662	0.662	0.648	0.656	0.656	0.645
		CESAR4	0.879	0.878	0.879	0.877	0.897	0.769	0.877	0.865	0.868	0.868	0.868
		GRADE	0.774	0.775	0.775	0.770	0.778	0.493	0.765	0.518	0.656	0.481	0.479
		LIZARD	0.098	0.098	0.098	0.874	0.893	0.875	0.809	0.861	0.891	-	0.839
		VIRGIN	0.797	0.797	0.797	0.786	0.824	0.771	0.795	0.796	0.793	0.783	0.796
	polytom.	CHD	-	-	-	0.931	0.931	-	0.931	0.654	0.723	0.677	0.490
		COLLEGE	-	-	-	0.796	0.883	-	0.888	0.727	0.785	0.785	0.768
		CREDIT	-	-	-	0.763	0.755	0.687	0.765	0.683	0.765	0.772	0.649
		EDUC	-	-	-	0.710	0.710	-	0.709	0.358	0.425	0.426	0.298
		ESTEEM	-	-	-	0.697	0.697	0.710	0.697	0.519	0.569	0.573	0.435
		IRIS	-	-	-	0.872	0.918	0.665	0.852	0.829	0.756	0.639	0.853
		KRETSCHM	-	-	-	0.781	0.759	0.617	0.658	0.654	0.738	0.728	0.646
		VOTING	-	-	-	0.827	0.839	-	0.831	0.816	0.831	0.612	0.819
artif.	dichot.	DILLON	0.913	0.913	0.911	0.916	0.911	0.824	0.916	0.518	0.918	0.588	0.648
		MA435300	0.613	0.615	0.618	0.645	0.685	0.566	0.676	0.558	0.695	-	0.677
		MA435301	0.649	0.652	0.651	0.666	0.612	0.538	0.695	0.571	0.708	-	0.697
		MA435302	0.698	0.691	0.698	0.649	0.629	0.571	0.673	0.611	0.687	-	0.673
		MA435303	0.674	0.674	0.675	0.658	0.651	0.576	0.682	0.604	0.701	-	0.678
		MA435304	0.668	0.665	0.667	0.672	0.652	0.594	0.692	0.604	0.709	-	0.692
		MA435305	0.650	0.651	0.651	0.665	0.616	0.563	0.697	0.584	0.715	-	0.698
		MA435306	0.684	0.682	0.682	0.643	0.632	0.568	0.672	0.607	0.693	-	0.673
		MA435307	0.642	0.641	0.641	0.663	0.625	0.570	0.693	0.571	0.707	-	0.691

(CONTINUED)

Table A-3: Posterior based eta criterion

expectation of hold-out based estimates of eta			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435308	0.608	0.608	0.604	0.683	0.592	0.552	0.714	0.544	0.726	-	0.716
		MA435309	0.721	0.721	0.721	0.684	0.668	0.576	0.711	0.641	0.722	-	0.711
	polytom.	BANANA	-	-	-	0.959	0.820	0.800	0.964	0.818	-	0.838	0.964
		INTERAC1	-	-	-	0.929	0.616	0.655	0.932	0.623	0.932	0.932	0.932
		NORMAL01	-	-	-	0.773	0.777	-	0.694	0.724	0.767	0.769	0.679
		NORMAL02	-	-	-	0.672	0.708	-	0.676	0.582	0.726	0.716	0.672
		NORMAL03	-	-	-	0.973	0.984	-	0.971	0.933	0.978	0.979	0.971
		NORMAL11	-	-	-	0.862	0.969	0.758	0.796	0.857	0.887	0.888	0.789
		NORMAL12	-	-	-	0.889	0.976	0.784	0.888	0.897	0.938	0.936	0.882
		NORMAL13	-	-	-	0.941	0.988	0.889	0.942	0.931	0.961	0.962	0.939
		NORMAL14	-	-	-	0.954	0.966	0.824	0.956	0.944	0.964	0.965	0.957
		NORMAL15	-	-	-	0.946	0.957	0.855	0.945	0.937	0.954	0.955	0.946
		NORMAL16	-	-	-	0.944	0.953	0.873	0.947	0.941	0.952	0.952	0.946
		NORMAL17	-	-	-	0.944	0.973	0.865	0.944	0.848	0.944	0.943	0.943
		POISSON	-	-	-	0.933	0.948	-	0.938	0.929	0.934	0.934	0.929

Table A-3: Posterior based eta criterion

Appendix B - Variability of performance criteria

var. of resubstitution based performance averaged across direct procedures		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
type	data							
real	BREAST	1.4	0.005	0.6	0.027	1.9	0.013	7
	CESAR4	2.9	0.004	56.1	0.078	4.4	0.039	7
	CHD	0.0	0.000	0.4	0.000	0.0	0.000	3
	COLLEGE	2.2	0.004	5.1	0.010	0.4	0.003	3
	CREDIT	18.3	0.041	24.3	0.058	6.2	0.048	4
	EDUC	0.0	0.000	0.3	0.001	0.1	0.000	3
	ESTEEM	0.0	0.000	5.0	0.015	1.0	0.007	4
	GRADE	80.2	0.242	4.4	0.009	14.3	0.100	7
	IRIS	80.5	0.146	87.7	0.136	14.3	0.120	4
	KRETSCHM	26.0	0.060	29.5	0.074	8.7	0.066	4
	LIZARD	10.4	0.012	0.3	0.009	0.6	0.005	7
	VIRGIN	0.0	0.000	14.0	0.030	1.9	0.015	7
	VOTING	0.0	0.000	0.5	0.014	0.0	0.007	3
artif.	BANANA	82.2	0.103	70.3	0.074	10.0	0.000	4
	DILLON	64.3	0.071	24.6	0.023	5.2	0.047	7
	INTERAC1	94.6	0.269	85.3	0.150	24.1	0.106	4
	MA435300	15.0	0.057	11.4	0.037	5.2	0.034	7
	MA435301	22.3	0.091	15.2	0.045	7.6	0.049	7
	MA435302	24.0	0.095	22.0	0.065	9.2	0.061	7
	MA435303	25.1	0.090	20.2	0.061	10.3	0.060	7
	MA435304	17.9	0.066	12.4	0.037	5.0	0.030	7
	MA435305	20.4	0.079	13.1	0.040	7.1	0.046	7
	MA435306	16.0	0.064	23.9	0.072	0.2	0.054	7
	MA435307	21.4	0.005	12.6	0.039	7.1	0.046	7

(CONTINUED)

Table B-1: Variability of resub. performance

var. of resubstitution based performance averaged across direct procedures		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
type	data							
artif.	MA435308	24.2	0.101	12.6	0.040	9.5	0.060	7
	MA435309	19.9	0.071	28.0	0.073	8.3	0.058	7
	NORMAL01	17.3	0.032	15.2	0.028	3.7	0.030	3
	NORMAL02	2.8	0.007	8.3	0.022	2.0	0.014	3
	NORMAL03	1.8	0.000	1.5	0.000	0.0	0.000	3
	NORMAL11	142.7	0.097	162.1	0.126	12.0	0.112	4
	NORMAL12	120.2	0.064	158.6	0.121	9.9	0.092	4
	NORMAL13	118.1	0.043	159.0	0.129	9.1	0.086	4
	NORMAL14	58.7	0.023	110.4	0.070	4.9	0.047	4
	NORMAL15	29.6	0.014	97.1	0.073	4.6	0.043	4
	NORMAL16	25.3	0.013	81.2	0.063	4.1	0.038	4
	NORMAL17	82.4	0.077	68.3	0.031	5.3	0.049	4
	POISSON	3.1	0.002	11.6	0.007	0.3	0.002	3

Table B-1: Variability of resub. performance

variability of resubstitution based performance averaged		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
procedure class	discriminant							
direct	bh1	48.8	8.138	31.3	8.871	13.5	8.898	16
	bh2	48.8	8.138	31.3	8.871	13.5	8.898	16
	bh3	48.8	8.138	31.3	8.871	13.5	8.898	16
	ker	66.1	8.113	67.6	8.126	14.5	8.119	37
	ldf	65.7	8.143	78.3	8.148	17.8	8.141	37
	lg1	58.4	8.283	33.3	8.898	18.1	8.125	29
	mlt	65.9	8.112	65.9	8.112	13.4	8.111	37
indirect	cen	62.8	8.176	61.4	8.182	23.8	8.171	37
	dd1	71.6	8.146	59.6	8.181	17.8	8.142	36
	dd2	91.5	8.281	74.6	8.896	24.6	8.192	26
	dhl	88.1	8.175	64.5	8.113	21.1	8.165	37

Table B-2: Variability of resub. performance

var. of hold-out based performance averaged across direct procedures		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
type	data							
real	BREAST	1.4	0.005	8.3	0.026	2.1	0.014	7
	CESAR4	6.7	0.009	57.2	0.088	5.0	0.043	7
	CHD	0.4	0.000	0.3	0.000	0.0	0.000	3
	COLLEGE	1.7	0.003	5.3	0.011	0.5	0.004	3
	CREDIT	7.0	0.019	35.3	0.086	5.0	0.037	4
	EDUC	0.0	0.000	0.4	0.001	0.1	0.000	3
	ESTEEM	0.0	0.000	4.2	0.013	0.9	0.006	4
	GRADE	66.2	0.226	14.0	0.031	14.4	0.106	7
	IRIS	55.0	0.115	85.5	0.144	13.2	0.109	4
	KRETSCHM	13.2	0.048	33.6	0.094	9.1	0.062	4
	LIZARD	9.3	0.011	6.9	0.000	0.9	0.000	7
	VIRGIN	2.4	0.005	14.7	0.029	2.0	0.016	7
	VOTING	1.5	0.003	9.0	0.015	0.7	0.006	3
artif.	BANANA	80.0	0.097	72.1	0.079	9.9	0.008	4
	DILLON	41.4	0.043	27.0	0.026	3.0	0.034	7
	INTERAC1	85.0	0.215	82.4	0.151	21.0	0.171	4
	MA435300	10.7	0.046	9.1	0.030	5.5	0.034	7
	MA435301	15.5	0.066	16.9	0.050	7.9	0.050	7
	MA435302	7.7	0.031	23.9	0.068	6.0	0.045	7
	MA435303	8.3	0.033	17.4	0.051	5.6	0.037	7
	MA435304	9.5	0.037	13.7	0.041	4.7	0.031	7
	MA435305	12.0	0.050	15.6	0.047	6.6	0.042	7
	MA435306	6.1	0.025	22.5	0.065	6.5	0.042	7
	MA435307	12.0	0.049	12.6	0.039	5.9	0.037	7

(CONTINUED)

Table B-3: Variability of hold-out performance

var. of hold-out based performance averaged across direct procedures		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
type	data							
artif.	MA435388	18.1	0.078	12.9	0.041	9.0	0.056	7
	MA435389	10.6	0.040	29.6	0.075	7.7	0.053	7
	NORMAL01	24.4	0.072	17.1	0.035	6.3	0.047	3
	NORMAL02	14.2	0.052	13.0	0.039	2.9	0.020	3
	NORMAL03	34.3	0.000	37.6	0.010	0.7	0.007	3
	NORMAL11	89.2	0.083	71.0	0.156	10.9	0.092	4
	NORMAL12	61.7	0.040	87.1	0.150	8.9	0.079	4
	NORMAL13	60.2	0.022	113.0	0.144	8.2	0.075	4
	NORMAL14	37.4	0.019	114.1	0.116	7.3	0.060	4
	NORMAL15	0.2	0.004	100.0	0.094	5.1	0.047	4
	NORMAL16	13.4	0.007	78.3	0.070	4.1	0.030	4
	NORMAL17	70.6	0.072	64.7	0.030	5.0	0.046	4
	POISSON	7.2	0.005	0.6	0.005	0.6	0.005	3

Table B-3: Variability of hold-out performance

variability of hold-out based performance averaged across data sets		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
procedure class	discriminant							
direct	bh1	39.7	0.132	31.3	0.071	13.8	0.099	16
	bh2	39.6	0.131	31.6	0.072	13.8	0.099	16
	bh3	39.5	0.131	31.5	0.071	13.8	0.099	16
	ker	66.0	0.139	54.5	0.108	15.0	0.119	37
	ldf	64.9	0.147	60.9	0.133	17.5	0.139	37
	lg1	61.9	0.200	31.0	0.095	17.6	0.122	29
	mlt	59.5	0.133	55.5	0.098	14.2	0.114	37
indirect	cen	59.1	0.172	61.3	0.102	22.3	0.158	37
	dd1	60.4	0.139	61.1	0.102	16.3	0.129	36
	dd2	85.5	0.178	76.7	0.096	21.0	0.164	26
	dhl	65.1	0.181	62.6	0.109	21.7	0.162	37

Table B-4: Variability of hold-out performance

var. of leave-1-out based performance averaged across direct procedures		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
type	data							
real	BREAST	1.8	0.006	8.5	0.026	2.8	0.014	7
	CESAR4	2.8	0.003	56.4	0.078	4.5	0.039	7
	CHD	0.0	0.000	0.5	0.000	0.0	0.000	3
	COLLEGE	2.1	0.004	5.1	0.010	0.5	0.004	3
	CREDIT	4.5	0.011	22.8	0.055	2.9	0.022	4
	EDUC	0.0	0.000	0.1	0.000	0.1	0.000	3
	ESTEEM	0.0	0.000	4.8	0.014	1.1	0.007	4
	GRADE	74.5	0.233	18.3	0.038	13.8	0.103	7
	IRIS	58.1	0.106	72.0	0.116	11.1	0.093	4
	KRETSCHM	6.8	0.020	26.5	0.070	5.2	0.037	4
	LIZARD	10.4	0.012	9.0	0.009	0.9	0.000	7
	VIRGIN	12.3	0.027	14.5	0.029	3.2	0.025	7
	VOTING	0.0	0.000	0.4	0.014	0.0	0.007	3
artif.	BANANA	82.1	0.102	73.9	0.000	10.2	0.090	4
	DILLON	63.4	0.075	24.3	0.023	5.5	0.049	7
	INTERAC1	94.6	0.269	85.5	0.151	24.1	0.106	4
	MA435300	6.6	0.027	12.1	0.039	4.5	0.028	7
	MA435301	22.1	0.095	15.5	0.045	0.4	0.053	7
	MA435302	10.9	0.079	23.0	0.064	9.4	0.060	7
	MA435303	21.5	0.085	10.0	0.054	9.4	0.061	7
	MA435304	14.7	0.056	11.0	0.034	5.2	0.034	7
	MA435305	19.2	0.000	13.6	0.040	0.2	0.053	7
	MA435306	12.4	0.040	24.6	0.073	7.2	0.047	7
	MA435307	15.7	0.065	13.7	0.042	6.4	0.040	7

(CONTINUED)

Table B-5: Variability of leave-1-out perf.

var. of leave-1-out based performance averaged across direct procedures		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
type	data							
artif.	MA435388	19.7	0.086	13.7	0.043	9.3	0.057	7
	MA435389	15.9	0.060	28.0	0.070	8.9	0.060	7
	NORMAL01	10.1	0.025	14.2	0.020	2.9	0.022	3
	NORMAL02	21.7	0.076	11.2	0.029	5.3	0.036	3
	NORMAL03	14.4	0.002	10.4	0.003	0.2	0.002	3
	NORMAL11	81.7	0.048	70.2	0.101	6.0	0.055	4
	NORMAL12	46.4	0.010	92.0	0.106	6.3	0.050	4
	NORMAL13	49.7	0.012	120.1	0.119	2.3	0.022	4
	NORMAL14	25.3	0.009	106.4	0.070	4.6	0.044	4
	NORMAL15	6.1	0.002	92.3	0.072	3.9	0.036	4
	NORMAL16	8.2	0.004	77.7	0.062	3.4	0.032	4
	NORMAL17	81.1	0.075	67.2	0.030	4.0	0.045	4
	POISSON	5.4	0.004	4.0	0.003	0.6	0.006	3

Table B-5: Variability of leave-1-out perf.

variability of leave-1-out based performance averaged across data sets		err_c		err_p		eta		count
		cv(%)	std	cv(%)	std	cv(%)	std	n
procedure class	discriminant							
direct	bh1	39.7	0.130	31.6	0.070	13.8	0.100	16
	bh2	39.7	0.130	31.6	0.070	13.8	0.100	16
	bh3	39.7	0.130	31.6	0.070	13.8	0.100	16
	ker	68.3	0.138	59.4	0.114	16.3	0.131	37
	ldf	66.8	0.151	70.0	0.130	18.4	0.145	37
	lg1	67.0	0.222	32.5	0.087	19.8	0.139	29
	mlt	61.4	0.123	59.7	0.103	14.2	0.115	37
indirect	cen	60.2	0.175	61.4	0.102	24.0	0.171	37
	dd1	69.6	0.141	61.0	0.102	17.0	0.136	36
	dd2	91.7	0.203	76.7	0.096	24.7	0.193	26
	dhl	77.4	0.184	64.4	0.113	22.1	0.171	37

Table B-6: Variability of leave-1-out perf.

Appendix C - Standard errors of performance criteria

standard error of hold-out based estimates of unconditional err(counting)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dh1
type	pred	data											
real	dichot.	BREAST	0.002	0.002	0.002	0.005	0.002	0.018	0.006	0.006	0.005	0.005	0.006
		CESAR4	0.002	0.002	0.002	0.003	0.004	0.172	0.003	0.003	0.003	0.003	0.003
		GRADE	0.010	0.011	0.010	0.014	0.023	0.018	0.024	0.025	0.024	0.017	0.025
		LIZARD	0.005	0.005	0.006	0.015	0.014	0.006	0.008	0.007	0.010	-	0.009
		VIRGIN	0.002	0.002	0.002	0.002	0.002	0.016	0.002	0.002	0.002	0.002	0.002
	polytom.	CHD	-	-	-	0.000	0.000	-	0.000	0.000	0.000	0.000	0.000
		COLLEGE	-	-	-	0.001	0.001	-	0.001	0.001	0.001	0.001	0.001
		CREDIT	-	-	-	0.006	0.004	0.016	0.007	0.008	0.008	0.007	0.008
		EDUC	-	-	-	0.000	0.000	-	0.001	0.001	0.001	0.000	0.002
		ESTEEM	-	-	-	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
		IRIS	-	-	-	0.011	0.006	0.064	0.017	0.018	0.014	0.016	0.015
		KRETSCHM	-	-	-	0.036	0.035	0.068	0.043	0.045	0.052	0.042	0.043
		VOTING	-	-	-	0.001	0.003	-	0.001	0.001	0.001	0.001	0.002
artif.	dichot.	DILLON	0.000	0.000	0.000	0.000	0.015	0.093	0.001	0.001	0.001	0.001	0.001
		MA435300	0.025	0.025	0.023	0.018	0.020	0.051	0.013	0.013	0.016	-	0.012
		MA435301	0.000	0.006	0.009	0.013	0.026	0.060	0.009	0.013	0.011	-	0.014
		MA435302	0.008	0.009	0.009	0.015	0.013	0.001	0.012	0.013	0.011	-	0.012
		MA435303	0.014	0.014	0.015	0.017	0.025	0.000	0.016	0.013	0.014	-	0.013
		MA435304	0.019	0.019	0.020	0.016	0.019	0.077	0.019	0.018	0.020	-	0.018
		MA435305	0.013	0.012	0.016	0.018	0.029	0.056	0.015	0.018	0.017	-	0.019
		MA435306	0.006	0.006	0.006	0.019	0.018	0.091	0.014	0.015	0.015	-	0.013
		MA435307	0.011	0.016	0.019	0.017	0.025	0.047	0.012	0.013	0.015	-	0.012

(CONTINUED)

Table C-1: Standard error of err (counting)

standard error of hold-out based estimates of unconditional err(counting)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435388	0.022	0.020	0.023	0.015	0.040	0.038	0.015	0.012	0.012	-	0.013
		MA435389	0.008	0.007	0.006	0.014	0.019	0.050	0.009	0.009	0.010	-	0.010
	polyton.	BANANA	-	-	-	0.001	0.043	0.022	0.000	0.001	-	0.000	0.000
		INTERAC1	-	-	-	0.000	0.000	0.106	0.000	0.000	0.000	0.000	0.000
		NORMAL01	-	-	-	0.016	0.009	-	0.021	0.018	0.021	0.022	0.020
		NORMAL02	-	-	-	0.029	0.024	-	0.031	0.033	0.031	0.029	0.029
		NORMAL03	-	-	-	0.002	0.001	-	0.002	0.002	0.002	0.002	0.002
		NORMAL11	-	-	-	0.007	0.008	0.109	0.019	0.022	0.018	0.020	0.020
		NORMAL12	-	-	-	0.007	0.003	0.021	0.013	0.013	0.014	0.013	0.015
		NORMAL13	-	-	-	0.006	0.003	0.005	0.000	0.011	0.010	0.010	0.010
		NORMAL14	-	-	-	0.005	0.004	0.194	0.006	0.006	0.006	0.007	0.007
		NORMAL15	-	-	-	0.006	0.004	0.027	0.007	0.007	0.006	0.007	0.007
		NORMAL16	-	-	-	0.007	0.003	0.004	0.007	0.007	0.006	0.007	0.006
		NORMAL17	-	-	-	0.002	0.002	0.039	0.002	0.002	0.007	0.002	0.007
		POISSON	-	-	-	0.002	0.002	-	0.003	0.003	0.002	0.003	0.002

Table C-1: Standard error of err (counting)

standard error of hold-out based estimates of unconditional err(posterior_1)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lgl	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.011	0.012	0.011	0.010	0.010	0.024	0.011	0.010	0.011	0.010	0.012
		CESAR1	0.004	0.004	0.004	0.004	0.013	0.095	0.005	0.006	0.005	0.004	0.005
		GRADE	0.019	0.018	0.018	0.019	0.026	0.039	0.022	0.023	0.026	0.019	0.021
		LIZARD	0.021	0.010	0.010	0.013	0.022	0.010	0.010	0.010	0.010	-	0.019
		VIRGIN	0.016	0.016	0.020	0.015	0.029	0.027	0.016	0.019	0.015	0.010	0.019
	polyton.	CHD	-	-	-	0.002	0.001	-	0.002	0.002	0.002	0.002	0.002
		COLLEGE	-	-	-	0.003	0.003	-	0.004	0.004	0.004	0.003	0.004
		CREDIT	-	-	-	0.010	0.006	0.012	0.010	0.010	0.009	0.011	0.010
		EDUC	-	-	-	0.001	0.001	-	0.001	0.001	0.001	0.001	0.001
		ESTEEM	-	-	-	0.001	0.000	0.009	0.001	0.001	0.001	0.001	0.001
		IRIS	-	-	-	0.016	0.023	0.043	0.020	0.020	0.010	0.020	0.020
		KRETSCHM	-	-	-	0.046	0.056	0.057	0.049	0.055	0.050	0.051	0.045
		VOTING	-	-	-	0.007	0.006	-	0.006	0.008	0.007	0.008	0.007
artif.	dichot.	DILLON	0.007	0.007	0.006	0.004	0.006	0.026	0.009	0.009	0.009	0.010	0.009
		MA435300	0.053	0.049	0.051	0.021	0.033	0.032	0.027	0.029	0.026	-	0.023
		MA435301	0.051	0.053	0.054	0.023	0.031	0.062	0.022	0.027	0.023	-	0.023
		MA435302	0.049	0.042	0.045	0.025	0.033	0.043	0.024	0.026	0.030	-	0.027
		MA435303	0.041	0.042	0.049	0.026	0.030	0.051	0.027	0.025	0.026	-	0.026
		MA435304	0.040	0.050	0.050	0.029	0.037	0.034	0.020	0.031	0.027	-	0.025
		MA435305	0.041	0.042	0.050	0.023	0.034	0.033	0.022	0.020	0.025	-	0.024
		MA435306	0.041	0.043	0.042	0.021	0.037	0.044	0.026	0.024	0.022	-	0.026
		MA435307	0.049	0.055	0.054	0.022	0.030	0.039	0.024	0.027	0.025	-	0.025

(CONTINUED)

Table C-2: Standard error of err (posterior)

standard error of hold-out based estimates of unconditional err(posterior_1)			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435308	0.049	0.045	0.046	0.021	0.035	0.043	0.024	0.023	0.026	-	0.021
		MA435309	0.034	0.038	0.035	0.023	0.034	0.031	0.021	0.024	0.020	-	0.026
	polytom.	BANANA	-	-	-	0.004	0.006	0.012	0.005	0.004	-	0.004	0.004
		INTERAC1	-	-	-	0.002	0.003	0.012	0.002	0.002	0.002	0.002	0.002
		NORMAL01	-	-	-	0.019	0.022	-	0.021	0.020	0.021	0.021	0.022
		NORMAL02	-	-	-	0.038	0.041	-	0.040	0.037	0.037	0.041	0.040
		NORMAL03	-	-	-	0.003	0.001	-	0.003	0.002	0.003	0.003	0.003
		NORMAL11	-	-	-	0.015	0.003	0.051	0.012	0.013	0.011	0.012	0.012
		NORMAL12	-	-	-	0.016	0.004	0.055	0.011	0.010	0.011	0.011	0.011
		NORMAL13	-	-	-	0.013	0.005	0.045	0.008	0.009	0.007	0.008	0.008
		NORMAL14	-	-	-	0.012	0.004	0.047	0.009	0.009	0.009	0.009	0.010
		NORMAL15	-	-	-	0.015	0.005	0.024	0.012	0.013	0.011	0.012	0.012
		NORMAL16	-	-	-	0.016	0.009	0.036	0.014	0.015	0.013	0.014	0.014
		NORMAL17	-	-	-	0.015	0.004	0.021	0.015	0.014	0.016	0.015	0.016
		POISSON	-	-	-	0.005	0.002	-	0.005	0.004	0.004	0.004	0.004

Table C-2: Standard error of err (posterior)

standard error of hold-out based estimates of unconditional eta			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lgl	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.006	0.006	0.006	0.006	0.005	0.014	0.007	0.007	0.006	0.006	0.007
		CESAR4	0.002	0.002	0.002	0.003	0.006	0.109	0.003	0.003	0.003	0.002	0.003
		GRADE	0.011	0.010	0.010	0.012	0.016	0.014	0.017	0.017	0.020	0.013	0.010
		LIZARD	0.011	0.010	0.010	0.012	0.013	0.009	0.011	0.010	0.013	-	0.012
		VIRGIN	0.008	0.008	0.010	0.008	0.014	0.017	0.008	0.010	0.007	0.009	0.010
	polytom.	CHD	-	-	-	0.001	0.001	-	0.001	0.001	0.001	0.001	0.001
		COLLEGE	-	-	-	0.002	0.001	-	0.002	0.002	0.002	0.002	0.002
		CREDIT	-	-	-	0.007	0.004	0.011	0.007	0.007	0.006	0.008	0.007
		EDUC	-	-	-	0.000	0.000	-	0.001	0.001	0.001	0.001	0.001
		ESTEEM	-	-	-	0.001	0.000	0.005	0.001	0.001	0.001	0.001	0.001
		IRIS	-	-	-	0.010	0.012	0.043	0.013	0.015	0.012	0.014	0.013
		KRETSCHM	-	-	-	0.030	0.030	0.047	0.036	0.044	0.043	0.040	0.035
		VOTING	-	-	-	0.004	0.003	-	0.003	0.004	0.004	0.004	0.004
artif.	dichot.	DILLON	0.004	0.004	0.003	0.002	0.008	0.045	0.004	0.004	0.005	0.005	0.004
		MA435300	0.026	0.022	0.025	0.015	0.023	0.031	0.015	0.018	0.016	-	0.015
		MA435301	0.025	0.027	0.026	0.013	0.021	0.035	0.013	0.016	0.014	-	0.014
		MA435302	0.024	0.023	0.023	0.014	0.017	0.046	0.014	0.015	0.017	-	0.016
		MA435303	0.020	0.020	0.023	0.016	0.023	0.060	0.018	0.014	0.015	-	0.016
		MA435304	0.023	0.022	0.022	0.018	0.022	0.042	0.019	0.020	0.019	-	0.016
		MA435305	0.021	0.022	0.025	0.015	0.021	0.029	0.016	0.019	0.017	-	0.018
		MA435306	0.021	0.022	0.021	0.015	0.022	0.053	0.016	0.015	0.015	-	0.016
		MA435307	0.023	0.025	0.023	0.013	0.019	0.026	0.015	0.016	0.015	-	0.015

(CONTINUED)

Table C-3: Standard error of eta

standard error of hold-out based estimates of unconditional eta			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435388	0.023	0.020	0.021	0.013	0.020	0.026	0.015	0.015	0.016	-	0.013
		MA435389	0.019	0.020	0.018	0.013	0.020	0.050	0.012	0.013	0.012	-	0.014
	polytom.	BANANA	-	-	-	0.002	0.023	0.013	0.002	0.002	-	0.002	0.002
		INTERAC1	-	-	-	0.001	0.002	0.053	0.001	0.001	0.001	0.001	0.001
		NORMAL01	-	-	-	0.014	0.013	-	0.018	0.015	0.017	0.019	0.016
		NORMAL02	-	-	-	0.023	0.025	-	0.029	0.029	0.028	0.029	0.026
		NORMAL03	-	-	-	0.002	0.001	-	0.002	0.002	0.002	0.002	0.002
		NORMAL11	-	-	-	0.008	0.004	0.066	0.014	0.016	0.014	0.015	0.014
		NORMAL12	-	-	-	0.008	0.003	0.031	0.010	0.010	0.010	0.010	0.011
		NORMAL13	-	-	-	0.000	0.002	0.057	0.007	0.000	0.000	0.000	0.000
		NORMAL14	-	-	-	0.006	0.003	0.116	0.007	0.006	0.006	0.007	0.007
		NORMAL15	-	-	-	0.008	0.003	0.020	0.008	0.008	0.007	0.008	0.008
		NORMAL16	-	-	-	0.009	0.005	0.019	0.008	0.009	0.006	0.009	0.009
		NORMAL17	-	-	-	0.008	0.002	0.014	0.007	0.007	0.010	0.008	0.010
		POISSON	-	-	-	0.003	0.001	-	0.003	0.003	0.003	0.003	0.003

Table C-3: Standard errors of eta

standard error of hold-out based unconditional performance averaged			err_c	err_p	eta	count
			se	se	se	n
type	pred	data				
real	dichot.	BREAST	0.005	0.013	0.007	7
		CESAR4	0.027	0.018	0.018	7
		GRADE	0.016	0.023	0.013	7
		LIZARD	0.008	0.018	0.011	7
		VIRGIN	0.004	0.020	0.010	7
	polytom.	CHD	0.000	0.002	0.001	3
		COLLEGE	0.001	0.003	0.002	3
		CREDIT	0.008	0.010	0.007	4
		EDUC	0.000	0.001	0.000	3
		ESTEEM	0.000	0.003	0.002	4
		IRIS	0.024	0.026	0.020	4
		KRETSCHM	0.045	0.052	0.038	4
		VOTING	0.002	0.006	0.003	3
artif.	dichot.	DILLON	0.016	0.009	0.010	7
		MA435300	0.026	0.038	0.022	7
		MA435301	0.019	0.042	0.023	7
		MA435302	0.021	0.037	0.023	7
		MA435303	0.027	0.039	0.026	7
		MA435304	0.027	0.039	0.024	7
		MA435305	0.023	0.035	0.021	7
		MA435306	0.023	0.036	0.024	7
		MA435307	0.021	0.039	0.021	7
		MA435308	0.025	0.038	0.021	7
		MA435309	0.022	0.031	0.022	7

(CONTINUED)

Table C-4: Stand. error of perf. criteria

standard error of hold-out based unconditional performance averaged			err_c	err_p	eta	count
			se	se	se	n
type	pred	data				
artif.	polytom.	BANANA	0.017	0.006	0.010	4
		INTERAC1	0.027	0.005	0.014	4
		NORMAL01	0.015	0.021	0.015	3
		NORMAL02	0.028	0.040	0.026	3
		NORMAL03	0.001	0.002	0.002	3
		NORMAL11	0.036	0.020	0.023	4
		NORMAL12	0.011	0.022	0.013	4
		NORMAL13	0.026	0.018	0.018	4
		NORMAL14	0.052	0.018	0.033	4
		NORMAL15	0.011	0.014	0.010	4
		NORMAL16	0.005	0.019	0.010	4
		NORMAL17	0.011	0.014	0.008	4
		POISSON	0.002	0.004	0.002	3

Table C-4: Stand. error of perf. criteria

standard error of hold-out based conditional performance averaged		err_c	err_p	eta	count
		se	se	se	n
procedure class	discriminant				
direct	bh1	0.094	0.224	0.065	16
	bh2	0.094	0.224	0.065	16
	bh3	0.094	0.224	0.065	16
	ker	0.050	0.236	0.094	37
	ldf	0.063	0.232	0.085	37
	lg1	0.094	0.212	0.060	29
	mlt	0.050	0.241	0.096	37
indirect	cen	0.049	0.241	0.096	37
	dd1	0.051	0.240	0.095	36
	dd2	0.037	0.254	0.109	26
	dh1	0.049	0.241	0.096	37

Table C-5: Stand. error of perf. criteria

Appendix D - Bias of conditional estimates

hold-out based bias relative to conditional estimate of err(counting) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dh1
type	pred	data											
real	dichot.	BREAST	-0.3	0.0	0.3	2.2	-0.3	1.6	2.7	4.8	2.8	1.8	7.8
		CESAR4	1.5	1.8	1.2	4.5	4.6	19.3	4.3	11.0	14.1	14.0	15.1
		GRADE	14.4	11.6	11.9	28.1	26.1	2.7	38.5	160.6	73.4	187.1	194.6
		LIZARD	4.2	2.0	2.4	10.2	3.2	-0.1	1.2	20.1	-0.7	-	52.9
		VIRGIN	2.2	-1.0	1.7	-0.0	4.5	7.9	3.5	0.5	-1.0	44.7	-0.6
	polytom.	CHD	-	-	-	-0.1	0.1	-	-0.7	402.6	301.2	365.9	647.0
		COLLEGE	-	-	-	0.4	0.3	-	2.1	37.0	7.0	0.3	17.0
		CREDIT	-	-	-	43.0	0.7	5.0	56.7	64.1	10.0	7.5	119.2
		EDUC	-	-	-	0.0	-0.0	-	0.0	120.9	90.1	90.4	154.1
		ESTEEM	-	-	-	-0.0	0.0	0.0	0.0	50.9	42.7	40.9	90.9
		IRIS	-	-	-	43.7	0.1	10.0	83.0	63.6	120.7	261.5	73.2
		KRETSCHM	-	-	-	06.6	53.3	57.3	110.0	00.0	0.0	3.3	116.5
		VOTING	-	-	-	-0.5	2.4	-	0.0	9.1	-1.1	129.9	0.7
artif.	dichot.	DILLON	1.0	1.4	1.9	3.3	7.7	-27.4	9.2	501.0	12.0	390.5	342.0
		MA435300	5.0	2.5	4.0	27.7	5.6	20.2	21.2	49.4	-0.5	-	22.2
		MA435301	0.6	1.0	2.4	21.6	-2.0	1.2	19.5	51.4	0.4	-	17.1
		MA435302	2.7	3.0	2.1	23.5	3.7	-20.0	20.6	26.9	-1.4	-	10.6
		MA435303	-2.2	-1.3	-0.0	22.3	7.9	-21.5	21.7	35.6	1.0	-	24.0
		MA435304	0.9	2.1	1.1	10.4	3.5	-7.3	16.9	39.6	-0.6	-	17.4
		MA435305	3.7	2.3	1.7	25.7	0.0	-0.5	24.7	52.0	0.5	-	24.7
		MA435306	1.6	3.2	0.2	33.5	0.2	-5.3	25.7	35.0	-1.3	-	25.0
		MA435307	0.1	-1.3	0.9	24.9	2.4	-6.0	10.4	49.9	2.9	-	21.0

(CONTINUED)

Table D-1: Relative bias of err(counting)

hold-out based bias relative to conditional estimate of err(counting) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435308	3.2	2.1	1.8	16.1	-3.5	8.5	21.4	73.6	-8.1	-	22.2
		MA435309	5.6	1.5	3.8	22.8	4.6	-12.2	24.7	33.1	-8.2	-	28.4
	polytom.	BANANA	-	-	-	1.4	-15.1	8.3	-8.5	482.8	-	347.7	8.8
		INTERAC1	-	-	-	1.8	8.1	-21.7	8.6	456.8	-8.4	8.1	8.1
		NORMAL01	-	-	-	74.6	5.7	-	121.8	36.8	8.6	-2.6	148.4
		NORMAL02	-	-	-	59.4	16.1	-	61.5	81.6	-1.7	-8.3	59.3
		NORMAL03	-	-	-	29.8	5.6	-	186.5	315.8	1.3	3.3	118.6
		NORMAL11	-	-	-	348.4	-9.5	176.3	1861	448.8	9.2	-4.1	1925
		NORMAL12	-	-	-	189.1	-11.5	88.2	692.1	484.3	-8.6	5.5	761.6
		NORMAL13	-	-	-	46.1	27.1	48.8	311.4	298.7	-1.3	8.7	318.1
		NORMAL14	-	-	-	46.8	-2.3	168.5	189.2	118.7	-4.8	-8.4	87.3
		NORMAL15	-	-	-	27.9	14.5	41.8	45.1	47.6	-3.7	3.4	62.5
		NORMAL16	-	-	-	26.9	-3.5	8.9	19.8	29.1	1.6	1.2	23.9
		NORMAL17	-	-	-	1.5	-5.4	-3.8	-1.2	185.5	-8.2	6.2	8.2
		POISSON	-	-	-	19.8	3.8	-	24.9	11.4	-8.5	1.3	26.7

Table D-1: Relative bias of err(counting)

hold-out based bias relative to conditional estimate of err(posterior_1) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dh1
type	pred	data											
real	dichot.	BREAST	0.3	0.1	0.1	-0.6	-0.1	1.4	-0.7	-1.0	0.9	-1.5	-0.6
		CESAR4	0.0	0.3	0.5	-0.9	-1.9	0.7	-2.6	1.6	-5.9	-2.4	3.6
		GRADE	-2.2	-1.5	-2.7	0.0	1.2	37.7	-3.6	-6.7	12.9	-19.2	18.3
		LIZARD	1.5	0.4	-2.7	0.6	3.5	9.7	-1.2	9.1	-14.6	-	-3.6
		VIRGIN	-0.4	0.2	-1.7	1.1	-2.2	-3.1	-1.9	8.1	-12.4	6.7	15.6
	polytom.	CHD	-	-	-	-0.4	-0.4	-	0.0	-1.6	2.3	-0.9	-2.2
		COLLEGE	-	-	-	0.1	-0.4	-	-0.0	-0.1	0.3	0.7	-3.1
		CREDIT	-	-	-	-6.1	-1.6	14.5	-5.0	-6.9	-18.8	-5.3	-6.5
		EDUC	-	-	-	0.0	0.2	-	0.2	0.5	0.0	-0.3	49.7
		ESTEEM	-	-	-	-0.0	0.0	1.5	-0.0	0.2	0.2	-0.6	-0.7
		IRIS	-	-	-	-4.7	7.2	9.3	32.8	24.6	-17.7	-1.6	-4.8
		KRETSCHM	-	-	-	1.0	-7.4	14.0	36.3	-32.6	-16.9	-3.9	22.9
		VOTING	-	-	-	0.7	-0.3	-	0.4	1.1	3.3	5.9	5.1
artif.	dichot.	DILLON	0.1	-2.7	1.7	4.2	3.1	2.9	-0.3	-12.3	-0.9	11.5	-2.2
		MA435300	1.1	1.2	-0.5	-6.3	-5.6	31.6	-7.0	1.6	-10.3	-	-11.4
		MA435301	3.0	-0.6	0.5	-6.6	-4.2	21.5	-0.0	-22.6	-2.2	-	-17.3
		MA435302	1.7	0.5	-0.3	-6.6	-5.1	13.0	-9.6	-7.6	-4.3	-	-3.9
		MA435303	1.0	2.1	2.4	-7.3	-4.6	-4.1	-5.1	-23.1	-6.7	-	-2.1
		MA435304	0.0	0.5	0.5	-4.0	-3.3	0.4	-2.5	-3.2	-2.7	-	-11.0
		MA435305	-1.7	-1.6	-1.3	-5.6	-6.5	14.0	-9.1	-0.3	-14.9	-	-0.1
		MA435306	-1.0	-0.9	0.2	-7.9	-4.7	-3.0	-7.1	-11.7	-6.3	-	-11.2
		MA435307	0.9	1.5	2.0	-6.9	-5.3	6.4	-9.0	-3.4	-7.2	-	-14.5

(CONTINUED)

Table D-2: Relative bias of err(posterior)

hold-out based bias relative to conditional estimate of err(posterior_1) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435308	-2.3	-1.0	-0.1	-5.3	-8.3	25.1	-8.2	-11.9	-12.1	-	2.5
		MA435309	-1.4	-0.3	0.2	-7.3	-5.9	5.5	-13.4	-5.1	-14.1	-	-17.3
	polytom.	BANANA	-	-	-	0.2	-0.5	10.0	0.5	10.7	-	-9.0	6.8
		INTERAC1	-	-	-	-0.3	-0.1	15.0	-0.3	-3.4	1.0	2.2	0.3
		NORMAL01	-	-	-	-3.6	-1.7	-	41.7	27.7	-2.9	-11.7	1.9
		NORMAL02	-	-	-	31.0	-4.1	-	-0.3	9.3	-10.0	-11.1	-16.3
		NORMAL03	-	-	-	138.3	0.0	-	66.5	15.7	29.7	43.5	37.9
		NORMAL11	-	-	-	2565	-2.0	49.3	1824	681.2	522.2	491.5	779.4
		NORMAL12	-	-	-	1298	-1.7	41.4	633.1	212.0	107.9	232.9	319.3
		NORMAL13	-	-	-	547.7	0.0	23.3	235.9	122.6	-31.0	0.7	-17.2
		NORMAL14	-	-	-	132.4	-3.5	62.5	52.0	37.6	19.5	-57.0	50.0
		NORMAL15	-	-	-	41.0	0.3	27.3	22.5	-0.3	58.0	-50.0	17.6
		NORMAL16	-	-	-	40.7	3.1	13.0	16.9	-7.0	-6.4	10.5	76.4
		NORMAL17	-	-	-	3.5	300.0	-1.5	0.4	-27.6	-16.2	20.3	-2.4
		POISSON	-	-	-	-6.2	1.4	-	-3.1	3.5	3.9	-15.7	13.2

Table D-2: Relative bias of err(posterior)

hold-out based bias relative to conditional estimate of eta (in %)			procedure class									
			direct							indirect		
			discriminant							discriminant		
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dh1
type	pred	data											
real	dichot.	BREAST	0.0	-0.0	-0.1	-0.4	0.1	-0.7	-0.5	-2.5	-1.4	-1.3	-3.1
		CESAR4	-0.1	-0.2	-0.1	-0.3	-0.2	-1.7	-0.1	-1.5	-2.0	-2.0	-2.1
		GRADE	-1.6	-1.3	-1.2	-3.4	-3.2	-5.2	-4.3	-35.6	-10.1	-10.6	-10.3
		LIZARD	-0.3	-0.1	0.0	-1.1	-0.4	-0.5	-0.0	-3.3	0.0	-	-5.7
		VIRGIN	-0.2	0.1	0.0	-0.1	-0.3	-0.6	-0.2	0.1	-0.1	-11.6	-0.0
	polytom.	CHD	-	-	-	0.0	0.0	-	0.0	-29.7	-22.4	-27.3	-47.3
		COLLEGE	-	-	-	-0.1	0.0	-	-0.2	-9.3	-2.0	-2.1	-4.1
		CREDIT	-	-	-	-4.9	0.1	-4.2	-5.9	-15.4	-5.9	-5.1	-20.1
		EDUC	-	-	-	0.0	-0.0	-	-0.0	-50.7	-40.1	-40.0	-57.9
		ESTEEM	-	-	-	0.0	0.0	-0.3	0.0	-25.6	-10.4	-17.0	-37.7
		IRIS	-	-	-	-2.2	-0.2	-3.5	-5.5	-7.0	-15.0	-29.0	-5.1
		KRETSCHM	-	-	-	-10.9	-4.5	-12.7	-10.5	-10.1	-7.0	-9.1	-19.0
		VOTING	-	-	-	-0.0	-0.2	-	-0.0	-1.0	0.1	-26.4	-1.6
artif.	dichot.	DILLON	-0.1	0.1	-0.2	-0.3	-0.5	4.5	-0.0	-44.4	-0.6	-36.7	-29.4
		MA435300	-2.2	-1.2	-1.5	-4.4	-0.0	-14.9	-3.1	-20.2	-0.7	-	-3.3
		MA435301	-1.0	-0.2	-0.9	-3.1	2.4	-6.4	-2.2	-19.6	-0.2	-	-1.4
		MA435302	-1.0	-0.9	-0.6	-3.7	0.3	6.7	-2.5	-11.2	0.0	-	-2.1
		MA435303	0.4	0.0	-0.2	-3.3	-0.9	14.1	-3.4	-13.9	-0.6	-	-4.0
		MA435304	-0.4	-0.7	-0.4	-2.9	-0.1	0.5	-2.9	-15.2	-0.5	-	-2.0
		MA435305	-0.0	-0.4	-0.3	-3.9	1.7	-3.9	-3.0	-10.0	-0.7	-	-3.2
		MA435306	-0.1	-0.0	-0.1	-5.4	1.3	3.0	-4.0	-13.6	-0.0	-	-4.0
		MA435307	-0.2	-0.5	-0.7	-3.5	0.9	0.9	-1.0	-19.4	-0.6	-	-2.9

(CONTINUED)

Table D-3: Relative bias of eta

hold-out based bias relative to conditional estimate of eta (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435308	-0.6	-0.5	-0.6	-2.0	4.4	-6.5	-2.4	-25.4	-0.1	-	-2.1
		MA435309	-1.2	-0.3	-0.8	-3.0	0.1	4.0	-2.1	-11.4	-0.4	-	-1.7
	polyton.	BANANA	-	-	-	-0.0	2.2	-2.0	0.0	-15.1	-	-13.1	0.0
		INTERAC1	-	-	-	-0.0	0.0	0.9	-0.0	-33.2	0.0	0.0	0.0
		NORMAL01	-	-	-	-6.9	-0.6	-	-16.6	-13.1	-0.1	-7.9	-19.0
		NORMAL02	-	-	-	-10.6	-2.1	-	-10.1	-22.7	-3.4	-4.6	-10.3
		NORMAL03	-	-	-	-1.2	-0.1	-	-1.4	-5.2	-0.6	-0.6	-1.4
		NORMAL11	-	-	-	-13.0	0.2	-10.9	-19.6	-13.4	-10.3	-10.4	-20.3
		NORMAL12	-	-	-	-9.9	0.2	-0.1	-9.9	-9.0	-4.0	-5.0	-10.5
		NORMAL13	-	-	-	-4.5	-0.3	-4.6	-4.3	-5.5	-2.4	-2.4	-4.6
		NORMAL14	-	-	-	-1.9	0.1	-0.5	-2.0	-3.3	-1.1	-1.1	-1.0
		NORMAL15	-	-	-	-1.4	-0.3	-3.7	-1.5	-2.4	-0.5	-0.5	-1.7
		NORMAL16	-	-	-	-1.0	0.0	-1.5	-0.9	-1.5	-0.4	-0.5	-0.9
		NORMAL17	-	-	-	-0.1	0.0	0.5	0.0	-10.9	-0.1	-0.4	-0.1
		POISSON	-	-	-	-0.5	-0.1	-	-0.7	-0.0	-0.2	-0.2	-0.0

Table D-3: Relative bias of eta

bias of hold-out based conditional performance averaged across direct			err_c	err_p	eta	count
			%	%	%	n
type	pred	data				
real	dichot.	BREAST	0.892	0.056	-0.236	7
		CESAR4	5.300	-0.542	-0.384	7
		GRADE	19.017	4.249	-2.882	7
		LIZARD	4.425	1.688	-0.363	7
		VIRGIN	2.576	-1.119	-0.183	7
	polytom.	CHD	-0.241	-0.298	0.018	3
		COLLEGE	0.936	-0.345	-0.075	3
		CREDIT	28.041	0.431	-3.728	4
		EDUC	-0.011	0.117	-0.014	3
		ESTEEM	-0.000	0.366	-0.070	4
		IRIS	34.381	11.178	-2.828	4
		KRETSCHM	76.818	11.195	-11.656	4
		VOTING	0.634	0.271	-0.096	3
artif.	dichot.	DILLON	-0.415	0.122	0.499	7
		MA435300	13.666	2.062	-3.899	7
		MA435301	6.224	0.808	-1.611	7
		MA435302	5.085	-0.790	-0.252	7
		MA435303	3.724	-2.109	0.946	7
		MA435304	5.084	0.042	-0.998	7
		MA435305	8.344	-1.571	-1.526	7
		MA435306	8.437	-3.717	-0.752	7
		MA435307	5.623	-1.170	-0.729	7
		MA435308	5.942	-0.001	-1.186	7
		MA435309	7.141	-3.215	-0.469	7

(CONTINUED)

Table D-4: Relative bias of perf. criteria

bias of hold-out based conditional performance averaged across direct			err_c	err_p	eta	count
			%	%	%	n
type	pred	data				
artif.	polytom.	BANANA	-1.478	2.586	0.020	4
		INTERAC1	-4.998	3.783	2.226	4
		NORMAL01	67.355	12.133	-8.023	3
		NORMAL02	45.665	9.148	-7.605	3
		NORMAL03	47.309	65.572	-0.863	3
		NORMAL11	591.935	1109.093	-10.803	4
		NORMAL12	239.447	492.652	-6.913	4
		NORMAL13	108.356	203.744	-3.442	4
		NORMAL14	78.370	60.868	-3.075	4
		NORMAL15	32.107	22.788	-1.726	4
		NORMAL16	13.038	18.412	-0.858	4
		NORMAL17	-2.244	75.578	0.109	4
		POISSON	15.898	-2.646	-0.445	3

Table D-4: Relative bias of perf. criteria

bias of hold-out based conditional performance averaged across data sets		err_c	err_p	eta	count
		%	%	%	n
procedure class	discriminant				
direct	bh1	2.813	0.150	-0.599	16
	bh2	1.929	0.065	-0.437	16
	bh3	2.280	-0.079	-0.475	16
	ker	36.076	127.396	-2.973	37
	ldf	4.047	6.666	-0.003	37
	lg1	18.663	15.480	-1.951	29
	mlt	104.784	77.155	-3.503	37
indirect	cen	130.449	25.996	-14.883	37
	dd1	19.203	14.925	-4.751	36
	dd2	74.298	24.738	-11.369	26
	dh1	150.521	34.402	-10.085	37

Table D-5: Relative bias of perf. criteria

Appendix E - Bias of unconditional estimates

hold-out based bias relative to unconditional estimate of err(counting) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lgi	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.2	0.4	0.4	1.6	0.4	0.8	1.5	4.1	1.8	1.3	6.6
		CESAR1	0.8	1.2	0.6	2.3	2.8	-15.5	2.2	9.2	11.8	10.9	12.3
		GRADE	10.0	8.9	11.2	19.1	10.6	1.8	32.8	132.7	65.5	159.3	170.8
		LIZARD	0.7	2.8	3.1	6.2	-1.4	0.7	1.6	27.5	-4.0	-	50.6
		VIRGIN	1.1	0.6	2.1	-0.3	1.8	5.8	2.5	0.0	-0.5	44.4	-0.7
	polytom.	CHD	-	-	-	-0.1	0.1	-	-0.7	402.6	301.2	365.9	647.0
		COLLEGE	-	-	-	0.5	0.4	-	1.5	37.1	7.5	7.6	17.3
		CREDIT	-	-	-	30.1	2.5	1.8	32.9	36.4	-6.2	-9.4	85.4
		EDUC	-	-	-	0.0	-0.0	-	-0.0	120.0	90.0	90.4	153.9
		ESTEEM	-	-	-	-0.0	0.0	0.0	0.0	50.9	42.7	40.9	90.0
		IRIS	-	-	-	26.6	3.1	6.6	30.5	30.9	85.0	174.2	37.6
		KRETSCHM	-	-	-	53.1	30.2	21.6	65.7	41.1	-24.3	-21.4	59.7
		VOTING	-	-	-	-1.0	1.5	-	0.1	9.2	-0.7	130.7	7.9
artif.	dichot.	DILLON	1.0	1.4	1.9	3.3	3.0	4.9	9.3	492.4	12.3	402.6	337.0
		MA435300	3.9	2.5	2.4	16.7	6.9	0.9	16.0	42.0	-6.1	-	15.0
		MA435301	0.7	1.7	1.0	17.4	4.2	2.6	15.2	44.6	-2.9	-	13.9
		MA435302	1.1	1.7	1.7	13.0	5.7	-4.0	14.9	21.6	-3.6	-	14.2
		MA435303	1.3	0.9	0.7	16.7	7.0	4.3	12.9	30.6	-2.9	-	16.3
		MA435304	0.7	3.3	2.1	12.2	4.0	-3.3	12.2	33.3	-5.2	-	12.9
		MA435305	2.9	1.3	2.5	15.7	7.0	-2.2	16.3	42.1	-6.0	-	14.3
		MA435306	1.3	1.9	1.2	19.7	6.2	-3.6	17.4	24.3	-7.2	-	17.0
		MA435307	1.2	-0.4	1.6	14.9	4.6	-2.4	15.7	43.6	-3.0	-	14.2

(CONTINUED)

Table E-1: Relative bias of err(counting)

hold-out based bias relative to unconditional estimate of err(counting) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435388	2.9	1.6	3.8	13.6	10.8	1.8	14.4	65.6	-2.6	-	15.2
		MA435389	1.8	1.8	1.5	17.8	8.4	-1.3	19.8	28.7	-3.4	-	16.6
	polytom.	BANANA	-	-	-	-8.3	9.5	5.8	-8.3	397.3	-	344.8	8.5
		INTERAC1	-	-	-	8.7	8.1	-14.8	8.1	455.1	-8.4	8.8	8.4
		NORMAL01	-	-	-	48.4	2.3	-	68.2	-1.7	-28.7	-29.1	67.4
		NORMAL02	-	-	-	35.2	9.4	-	34.4	51.9	-15.8	-14.8	37.8
		NORMAL03	-	-	-	14.8	1.8	-	59.6	289.8	-23.4	-21.2	59.8
		NORMAL11	-	-	-	129.2	7.5	22.6	194.8	-23.3	-84.1	-87.6	196.7
		NORMAL12	-	-	-	74.4	-1.9	56.8	164.8	43.8	-66.3	-68.8	168.8
		NORMAL13	-	-	-	15.1	10.6	-13.1	107.8	75.9	-45.5	-44.9	108.4
		NORMAL14	-	-	-	38.1	13.2	-23.9	62.2	71.1	-23.3	-20.9	52.3
		NORMAL15	-	-	-	21.6	9.3	8.8	31.9	28.7	-14.3	-11.8	31.7
		NORMAL16	-	-	-	24.1	-2.2	5.6	10.8	15.8	-9.8	-10.1	17.8
		NORMAL17	-	-	-	1.8	-5.4	10.8	3.6	198.2	-2.5	-1.1	8.8
		POISSON	-	-	-	12.1	1.2	-	17.1	3.8	-5.8	-4.5	17.9

Table E-1: Relative bias of err(counting)

hold-out based bias relative to unconditional estimate of err(posterior_1) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lgl	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	0.6	0.5	0.3	0.1	0.3	0.1	0.1	-0.3	2.0	-0.5	0.9
		CESAR1	0.0	0.0	0.0	-1.0	-0.1	2.7	-1.1	2.0	-4.7	-1.3	4.0
		GRADE	0.9	1.9	1.2	3.5	5.3	20.3	0.9	1.0	10.0	-14.9	26.1
		LIZARD	1.5	4.5	-0.3	1.0	-4.4	10.0	-0.4	0.5	-14.4	-	-1.0
		VIRGIN	-0.1	0.2	-0.9	4.0	-3.0	0.0	2.0	12.3	-0.9	0.6	10.0
	polyton.	CHD	-	-	-	-0.3	-1.0	-	-0.3	-1.7	2.7	-0.7	-2.2
		COLLEGE	-	-	-	0.4	-0.5	-	-0.3	0.4	0.0	1.3	-2.4
		CREDIT	-	-	-	1.3	-0.3	14.0	6.3	3.9	-0.3	6.6	5.3
		EDUC	-	-	-	0.0	0.2	-	0.1	0.5	0.0	-0.2	49.7
		ESTEEM	-	-	-	-0.1	0.0	-0.2	-0.0	0.2	0.2	-0.6	-0.7
		IRIS	-	-	-	1.2	-9.9	27.1	24.5	19.5	-20.0	-3.9	-0.2
		KRETSCHM	-	-	-	12.9	6.0	19.6	36.3	-31.3	-17.4	-1.0	17.2
		VOTING	-	-	-	0.6	0.3	-	0.2	0.9	3.0	6.1	4.7
artif.	dichot.	DILLON	-0.7	-1.9	3.0	2.6	2.1	0.7	-1.9	-6.4	7.2	20.3	4.4
		MA435300	0.0	2.3	0.3	-0.2	-0.2	19.6	1.6	9.3	-3.7	-	-2.1
		MA435301	4.9	-0.5	-5.1	-1.4	-0.6	14.0	2.0	-14.0	10.3	-	-7.0
		MA435302	2.9	-0.6	1.0	-2.7	-2.7	10.5	-3.3	2.3	4.4	-	5.5
		MA435303	3.4	1.4	1.1	-2.7	-1.3	9.0	-0.0	-10.2	-1.0	-	4.0
		MA435304	-0.5	0.5	-1.0	-1.7	-1.0	0.0	1.4	0.0	1.2	-	-0.3
		MA435305	-0.4	-0.7	-1.9	0.3	-2.2	15.0	-2.3	-1.9	-7.2	-	0.2
		MA435306	-2.1	-1.1	-1.3	-3.5	0.1	5.4	0.0	-3.0	1.4	-	-2.3
		MA435307	2.0	4.3	-4.6	-1.7	-0.3	12.4	0.2	7.4	1.2	-	-5.9

(CONTINUED)

Table E-2: Relative bias of err(posterior)

hold-out based bias relative to unconditional estimate of err(posterior_1) (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dhl
type	pred	data											
artif.	dichot.	MA435308	-2.1	-1.1	-1.0	-0.7	-2.8	14.1	-0.1	-3.4	-2.9	-	14.3
		MA435309	0.2	-1.0	-0.5	-0.8	0.2	9.0	-2.1	7.4	-2.1	-	-4.2
	polytom.	BANANA	-	-	-	-0.4	0.7	0.6	-0.5	10.4	-	-10.5	7.4
		INTERAC1	-	-	-	0.0	-0.1	9.0	0.0	-3.2	2.4	2.5	0.6
		NORMAL01	-	-	-	7.3	-2.1	-	37.7	23.9	-0.6	-17.3	-3.0
		NORMAL02	-	-	-	25.0	1.2	-	6.5	10.6	-3.5	-0.6	-7.9
		NORMAL03	-	-	-	66.7	0.6	-	44.9	1.7	16.2	25.6	10.5
		NORMAL11	-	-	-	209.0	0.0	20.0	199.0	15.4	-0.7	-16.0	32.5
		NORMAL12	-	-	-	211.9	-5.6	37.1	159.6	11.6	-20.3	21.5	56.9
		NORMAL13	-	-	-	159.4	5.4	27.9	09.7	24.0	-56.9	-42.9	-53.2
		NORMAL14	-	-	-	103.7	0.0	26.7	50.2	20.1	24.1	-50.7	43.9
		NORMAL15	-	-	-	23.3	0.6	31.0	22.0	-7.6	61.9	-40.6	10.2
		NORMAL16	-	-	-	32.3	-3.3	22.5	9.3	-13.3	-12.6	13.4	59.0
		NORMAL17	-	-	-	0.2	12.0	5.6	-1.0	-26.6	-19.2	15.3	-1.3
		POISSON	-	-	-	-1.3	0.4	-	4.1	10.0	11.1	-10.5	10.2

Table E-2: Relative bias of err(posterior)

hold-out based bias relative to unconditional estimate of eta (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lgl	mlt	cen	dd1	dd2	dhl
type	pred	data											
real	dichot.	BREAST	-0.2	-0.2	-0.2	-0.4	-0.2	-0.2	-0.4	-2.5	-1.4	-1.4	-3.2
		CESAR4	-0.1	-0.1	-0.1	-0.1	-0.2	1.2	-0.1	-1.5	-2.0	-1.9	-2.0
		GRADE	-1.7	-1.4	-1.7	-2.9	-2.0	-4.6	-4.3	-35.3	-10.2	-40.2	-40.1
		LIZARD	-0.1	-0.4	-0.2	-0.5	0.4	-0.6	-0.1	-3.3	0.2	-	-5.0
		VIRGIN	-0.1	-0.1	-0.2	-0.5	0.1	-0.9	-0.6	-0.4	-0.6	-11.7	-0.3
	polyton.	CHD	-	-	-	0.0	0.0	-	0.0	-29.7	-22.4	-27.3	-47.3
		COLLEGE	-	-	-	-0.1	0.0	-	-0.1	-9.3	-2.0	-2.1	-4.2
		CREDIT	-	-	-	-4.2	-0.4	-3.5	-5.1	-14.4	-5.1	-4.3	-19.5
		EDUC	-	-	-	0.0	-0.0	-	0.0	-58.7	-40.1	-40.0	-57.9
		ESTEEM	-	-	-	0.0	0.0	0.0	0.0	-25.6	-10.4	-17.0	-37.7
		IRIS	-	-	-	-1.0	0.1	-5.9	-3.0	-6.5	-14.0	-27.9	-3.0
		KRETSCHM	-	-	-	-9.0	-4.0	-9.5	-15.7	-15.4	-3.7	-5.7	-15.0
		VOTING	-	-	-	0.0	-0.2	-	-0.0	-1.0	-0.0	-26.4	-1.5
artif.	dichot.	DILLON	-0.0	0.0	-0.2	-0.3	-0.2	-1.3	-0.3	-44.5	-1.0	-37.0	-29.6
		MA435300	-1.6	-1.5	-0.9	-4.0	-2.1	-5.4	-4.1	-20.7	-0.9	-	-4.1
		MA435301	-1.2	-0.4	0.5	-3.6	-1.2	-5.3	-3.5	-20.4	-1.0	-	-2.9
		MA435302	-0.7	-0.4	-0.6	-2.0	-1.0	-1.5	-2.0	-12.2	-1.3	-	-3.2
		MA435303	-1.0	-0.5	-0.4	-3.3	-1.5	-4.2	-2.9	-14.3	-0.6	-	-4.0
		MA435304	-0.1	-1.1	-0.4	-2.3	-0.6	-1.2	-2.9	-15.0	-0.3	-	-2.6
		MA435305	-0.0	-0.3	-0.4	-3.6	-1.0	-3.4	-3.0	-10.0	-1.0	-	-3.0
		MA435306	0.0	-0.4	-0.1	-4.1	-1.0	-0.3	-4.1	-13.4	-1.0	-	-4.4
		MA435307	-1.0	-0.0	0.5	-3.0	-1.3	-2.4	-3.4	-20.4	-0.9	-	-3.4

(CONTINUED)

Table E-3: Relative bias of eta

hold-out based bias relative to unconditional estimate of eta (in %)			procedure class										
			direct							indirect			
			discriminant							discriminant			
			bh1	bh2	bh3	ker	ldf	lg1	mlt	cen	dd1	dd2	dh1
type	pred	data											
artif.	dichot.	MA435308	-0.5	-0.3	-0.9	-2.6	-2.7	-4.4	-2.8	-25.9	-1.3	-	-2.9
		MA435309	-0.5	-0.3	-0.3	-3.6	-2.2	-2.2	-3.4	-12.8	-2.8	-	-3.6
	polyton.	BANANA	-	-	-	0.0	-1.1	-1.5	0.0	-15.1	-	-13.8	0.0
		INTERAC1	-	-	-	-0.0	0.0	5.4	-0.0	-33.2	-0.0	0.0	-0.0
		NORMAL01	-	-	-	-5.7	-0.0	-	-12.9	-9.3	-3.6	-3.6	-15.0
		NORMAL02	-	-	-	-0.5	-2.0	-	-0.0	-21.2	-1.0	-3.5	-9.2
		NORMAL03	-	-	-	-0.0	-0.0	-	-1.0	-4.9	-0.3	-0.3	-1.0
		NORMAL11	-	-	-	-9.5	-0.1	-6.3	-14.5	-7.0	-4.5	-4.5	-15.3
		NORMAL12	-	-	-	-7.4	0.1	-7.3	-7.2	-6.1	-2.1	-2.1	-7.0
		NORMAL13	-	-	-	-3.2	-0.1	-4.0	-3.0	-4.1	-1.3	-1.2	-3.3
		NORMAL14	-	-	-	-1.0	-0.2	-2.0	-1.6	-2.9	-0.0	-0.7	-1.5
		NORMAL15	-	-	-	-1.0	-0.2	-3.1	-1.2	-2.1	-0.3	-0.2	-1.1
		NORMAL16	-	-	-	-1.0	0.1	-2.2	-0.5	-1.1	0.0	-0.1	-0.6
		NORMAL17	-	-	-	-0.1	0.1	-1.2	-0.1	-11.0	0.1	-0.1	-0.1
		POISSON	-	-	-	-0.4	-0.0	-	-0.7	-0.0	-0.2	-0.2	-0.7

Table E-3: Relative bias of eta

bias of hold-out based unconditional performance averaged across direct			err_c	err_p	eta	count
			%	%	%	n
type	pred	data				
real	dichot.	BREAST	0.760	0.295	-0.257	7
		CESAR4	-0.804	0.299	0.076	7
		GRADE	13.577	5.981	-2.650	7
		LIZARD	1.839	1.782	-0.215	7
		VIRGIN	1.938	0.296	-0.314	7
	polytom.	CHD	-0.241	-0.528	0.025	3
		COLLEGE	0.780	-0.099	-0.083	3
		CREDIT	16.824	5.339	-3.297	4
		EDUC	-0.021	0.117	-0.005	3
		ESTEEM	-0.008	-0.070	0.014	4
		IRIS	18.682	10.732	-2.842	4
		KRETSCHM	42.661	18.892	-9.753	4
		VOTING	0.195	0.376	-0.055	3
artif.	dichot.	DILLON	3.551	1.705	-0.338	7
		MA435300	7.174	3.480	-2.792	7
		MA435301	6.223	1.886	-2.089	7
		MA435302	4.982	0.738	-1.410	7
		MA435303	6.243	1.552	-1.987	7
		MA435304	4.463	0.806	-1.231	7
		MA435305	6.325	1.215	-1.903	7
		MA435306	6.304	-0.242	-1.551	7
		MA435307	5.024	1.873	-1.613	7
		MA435308	6.755	0.913	-2.030	7
		MA435309	6.988	0.730	-1.770	7

(CONTINUED)

Table E-4: Relative bias of perf. criteria

bias of hold-out based unconditional performance averaged across direct			err_c	err_p	eta	count
			%	%	%	n
type	pred	data				
artif.	polytom.	BANANA	3.482	2.075	-0.640	4
		INTERAC1	-3.452	2.427	1.336	4
		NORMAL01	34.306	14.281	-6.222	3
		NORMAL02	26.345	11.157	-6.186	3
		NORMAL03	25.115	37.416	-0.612	3
		NORMAL11	88.522	109.413	-7.633	4
		NORMAL12	73.353	100.734	-5.459	4
		NORMAL13	29.879	70.584	-2.592	4
		NORMAL14	22.406	45.140	-1.400	4
		NORMAL15	15.701	19.399	-1.404	4
		NORMAL16	9.366	15.221	-0.892	4
		NORMAL17	2.489	3.993	-0.307	4
		POISSON	10.141	1.040	-0.395	3

Table E-4: Relative bias of perf. criteria

bias of hold-out based unconditional performance averaged across data sets		err_c	err_p	eta	count
		%	%	%	n
procedure class	discriminant				
direct	bh1	2.011	0.755	-0.594	16
	bh2	1.935	0.611	-0.518	16
	bh3	2.353	-0.553	-0.343	16
	ker	19.094	22.968	-2.508	37
	ldf	4.751	-0.186	-0.715	37
	lg1	2.360	14.711	-2.691	29
	mlt	29.387	18.563	-3.082	37
indirect	cen	89.588	2.437	-14.443	37
	dd1	6.295	-0.986	-4.314	36
	dd2	55.244	-4.147	-10.512	26
	dh1	69.894	8.173	-9.682	37

Table E-5: Relative bias of perf. criteria

Appendix F - Performance over levels of discreteness

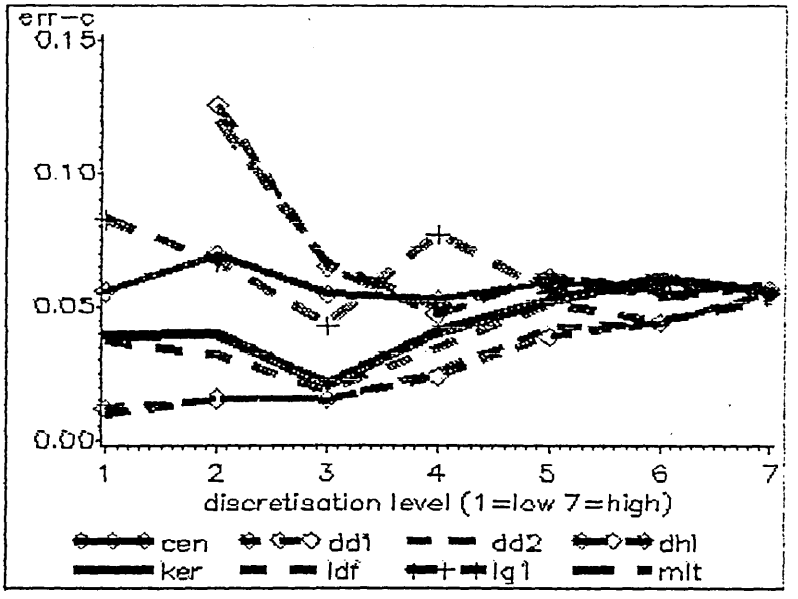


Figure F-1: Err(counting) leave-1-out estimate

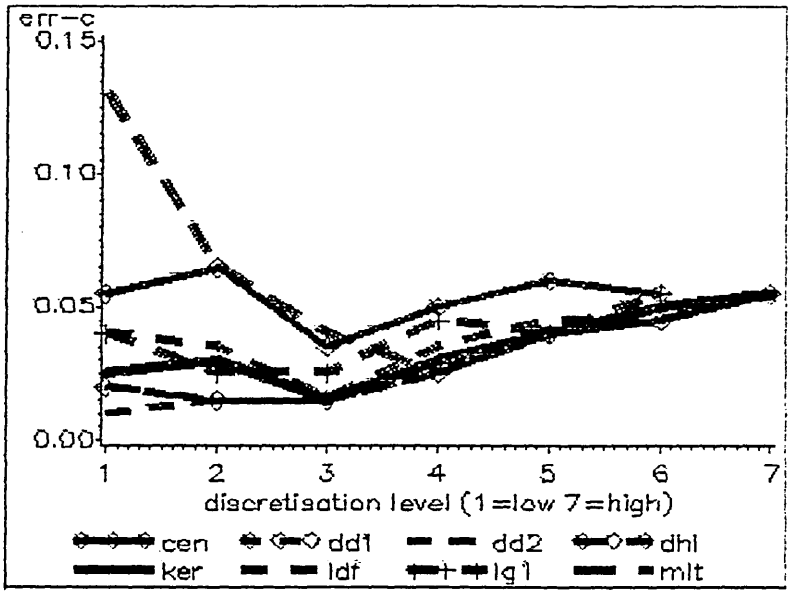


Figure F-2: Err(counting) hold-out estimate

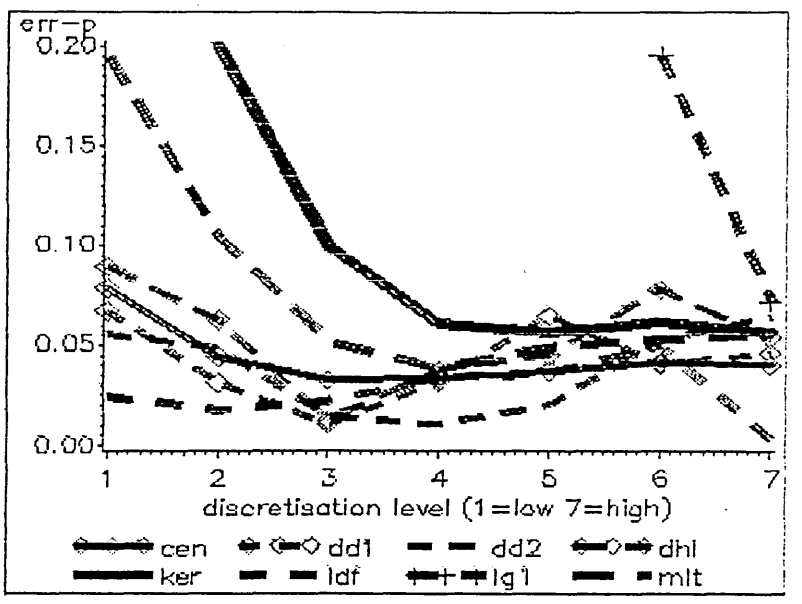


Figure F-3: Err(posterior) leave-1-out est.

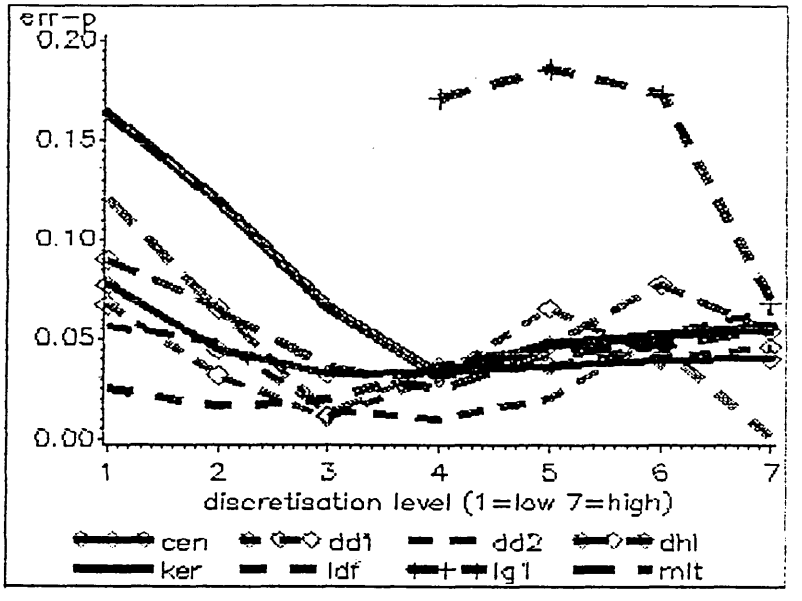


Figure F-4: Err(posterior) hold-out estimate

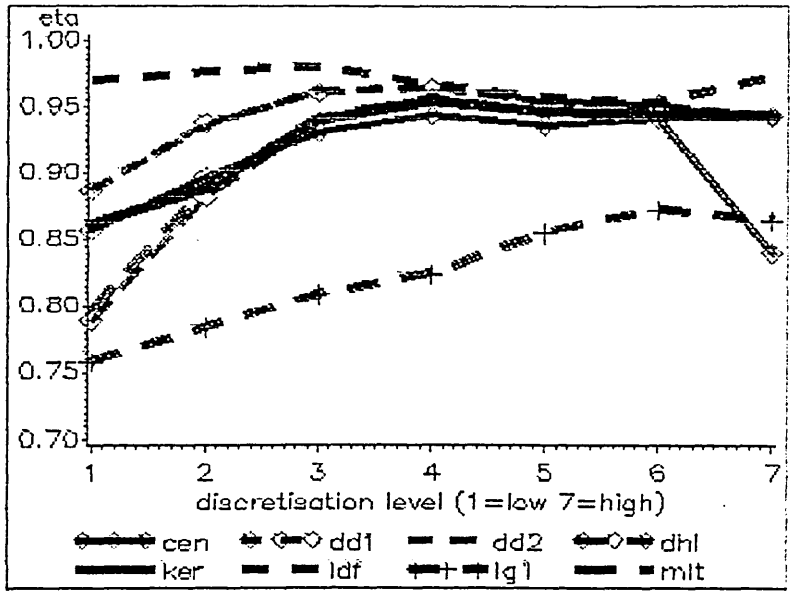


Figure F-5: η leave-one-out estimate

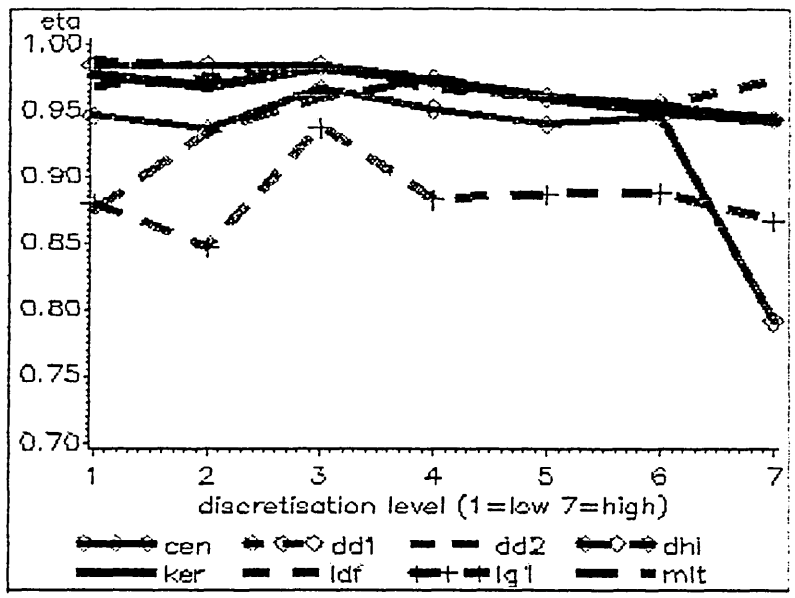


Figure F-6: η hold-out estimate

Appendix G - Performance over classification thresholds

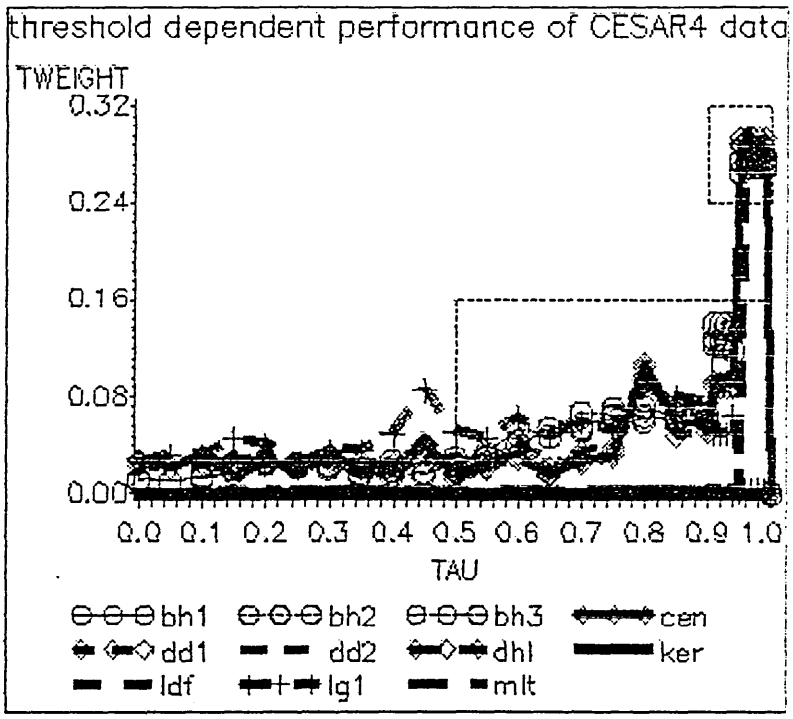


Figure G-1: Distribution of $f(\tau)$

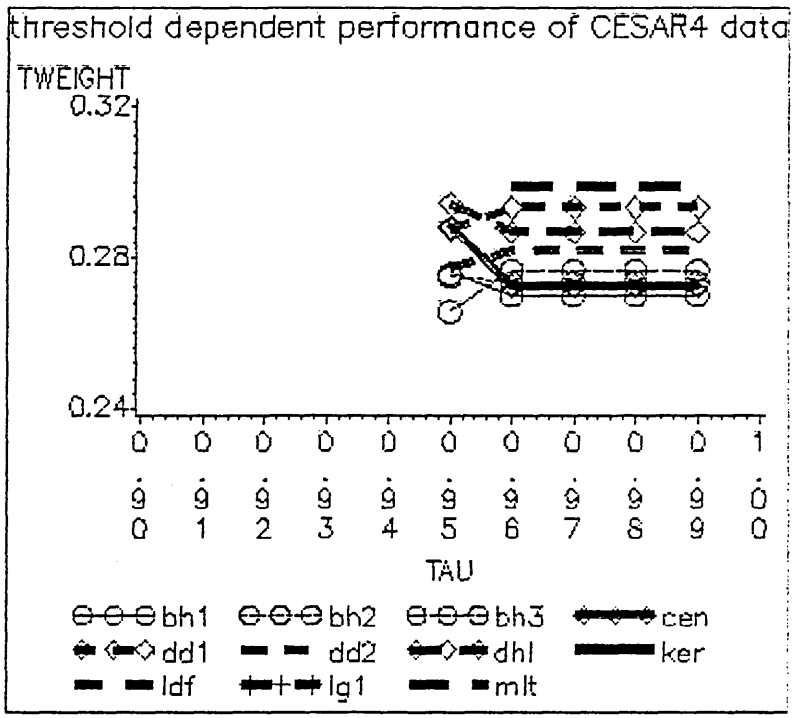


Figure G-2: Blow up of figure G-1

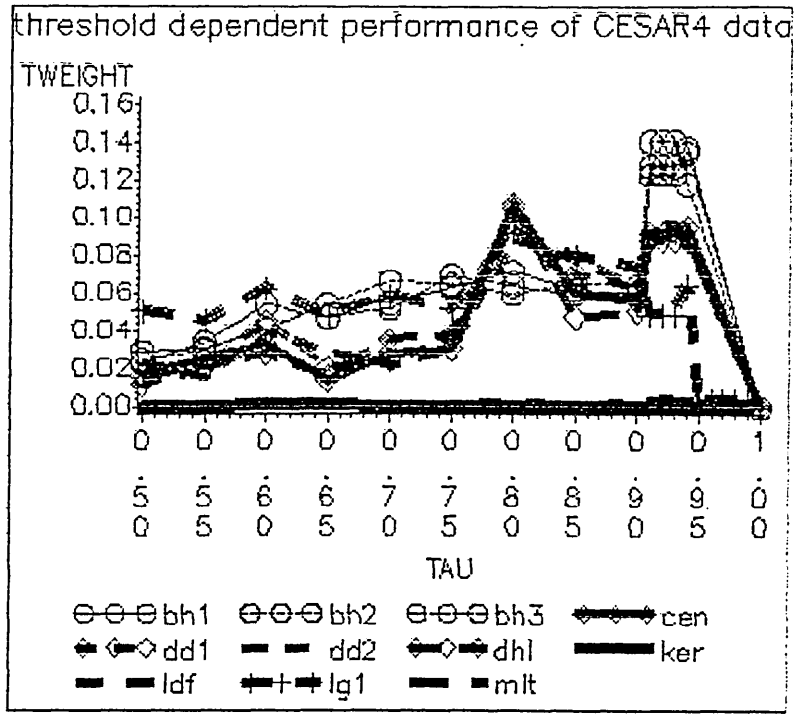


Figure G-3: Blow up of figure G-2

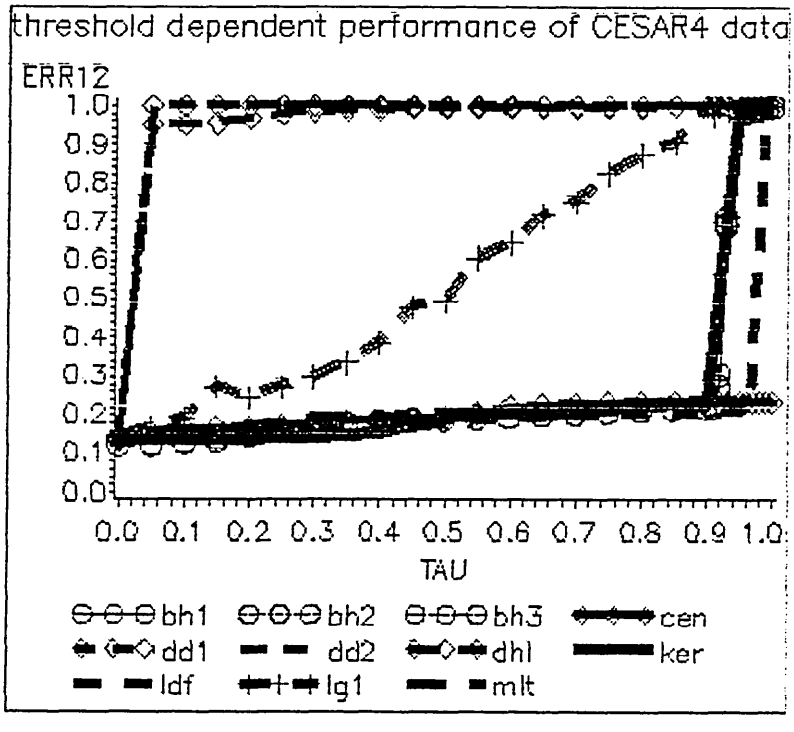


Figure G-4: Leave-1-out err(counting)

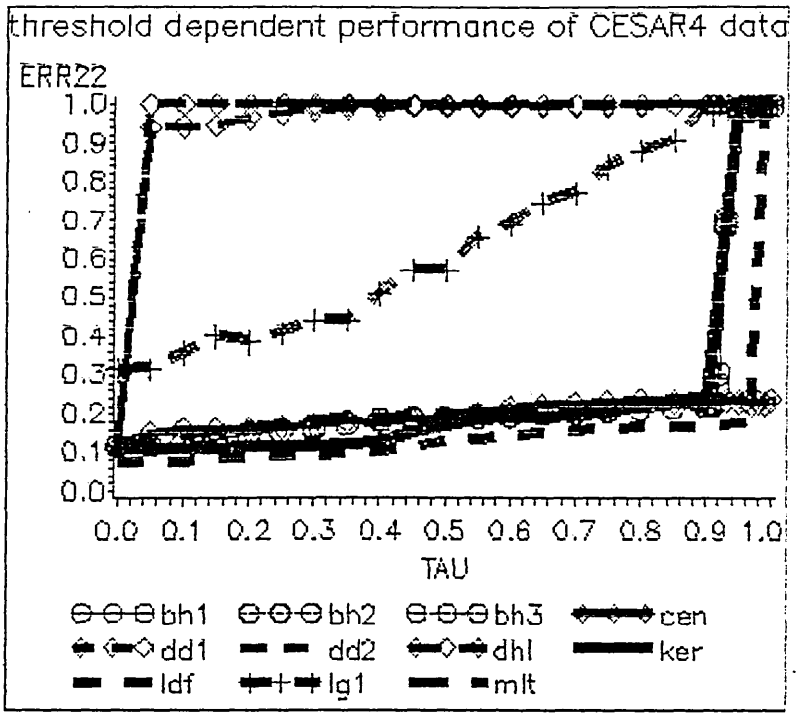


Figure G-5: Leave-1-out err(posterior)

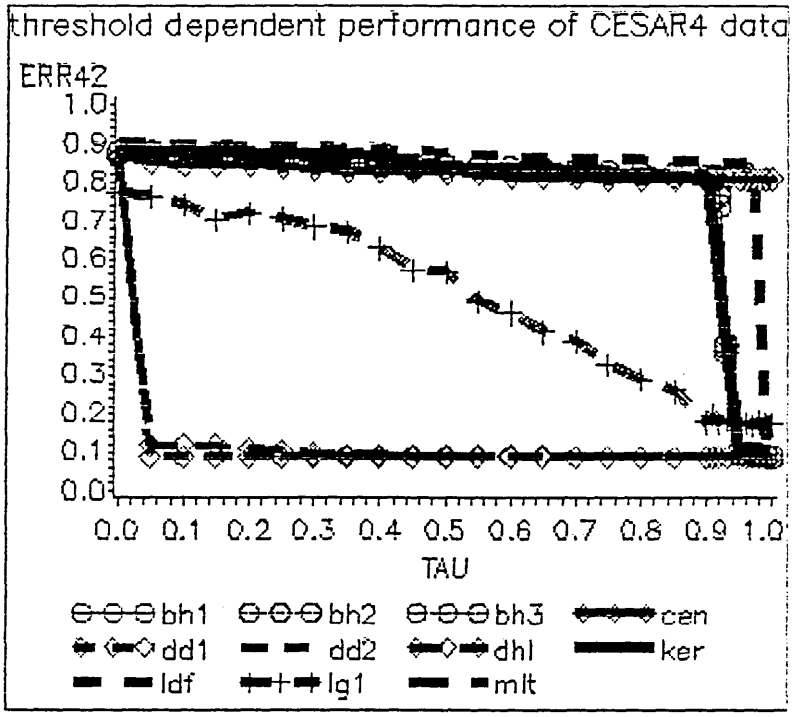


Figure G-6: Leave-1-out η

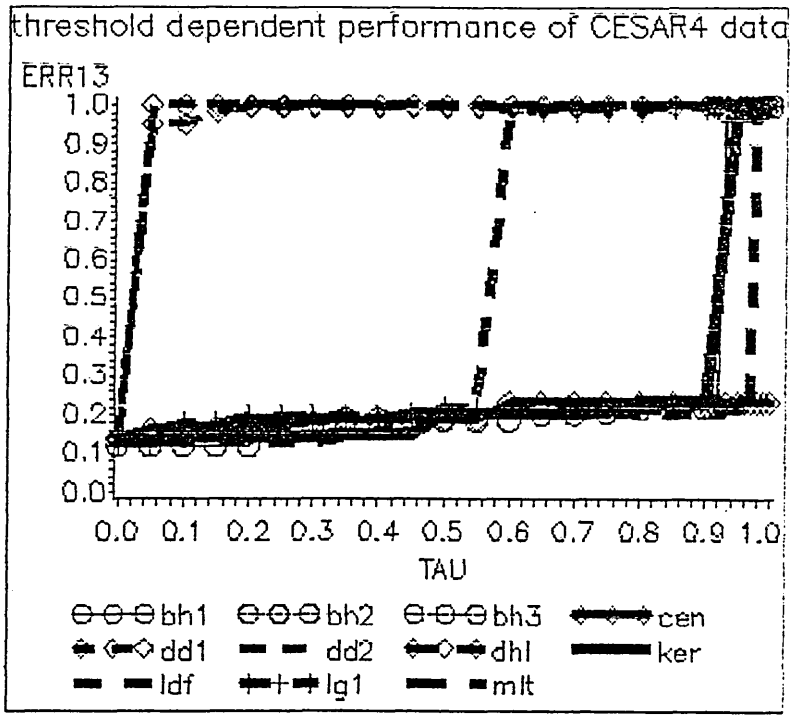


Figure G-7: Hold-out err(counting)

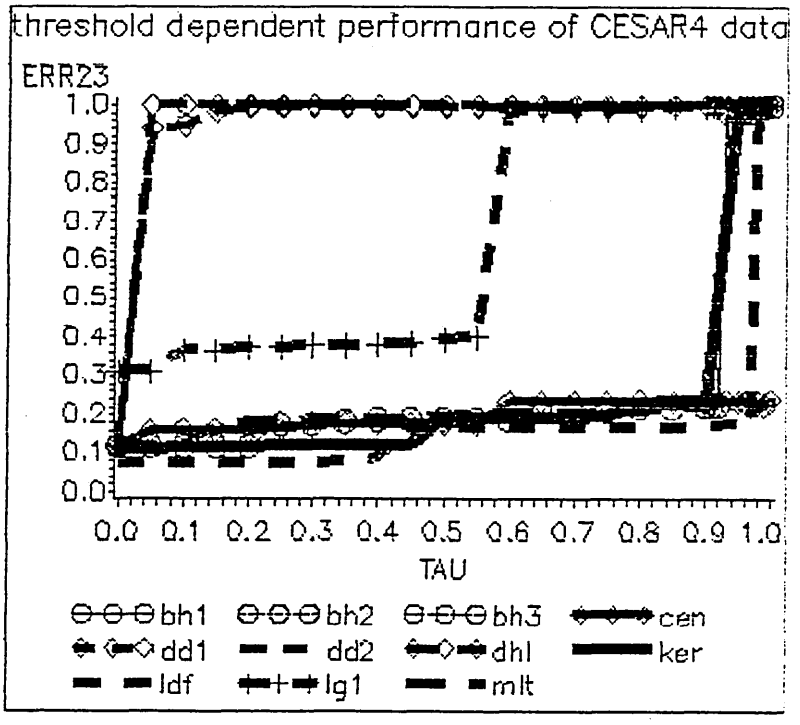


Figure G-8: Hold-out err(posterior)

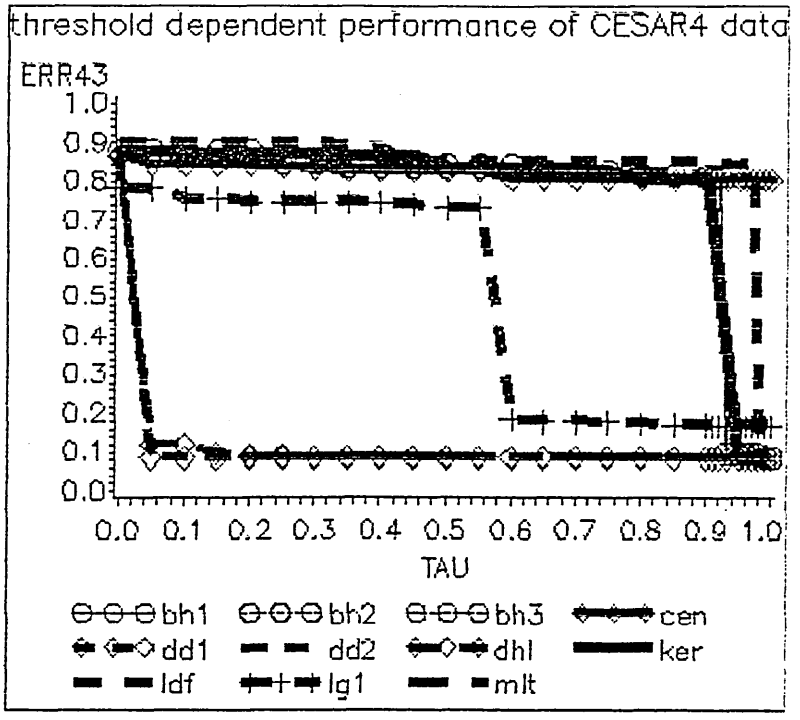


Figure G-9: Hold-out η

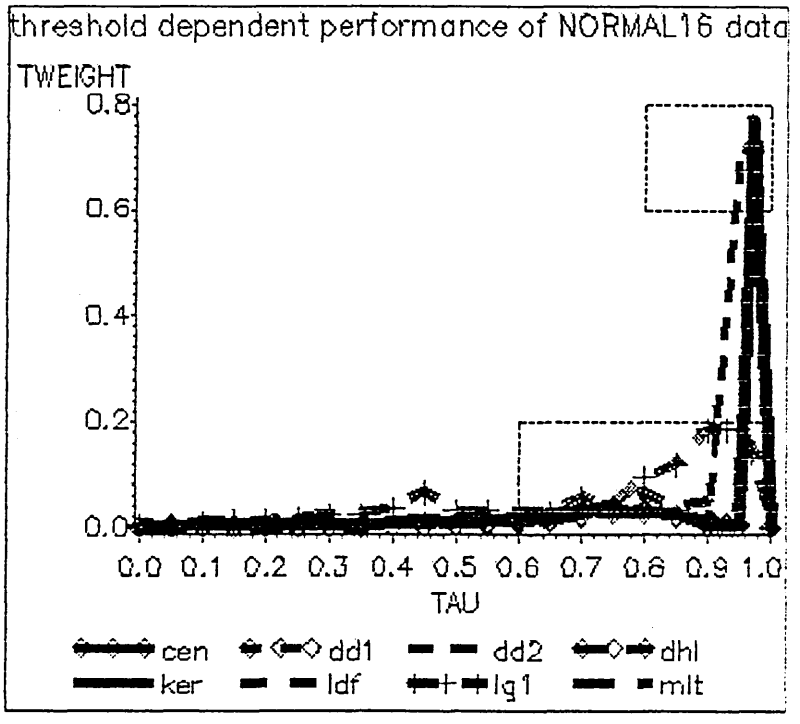


Figure G-10: Distribution of $f(\tau)$

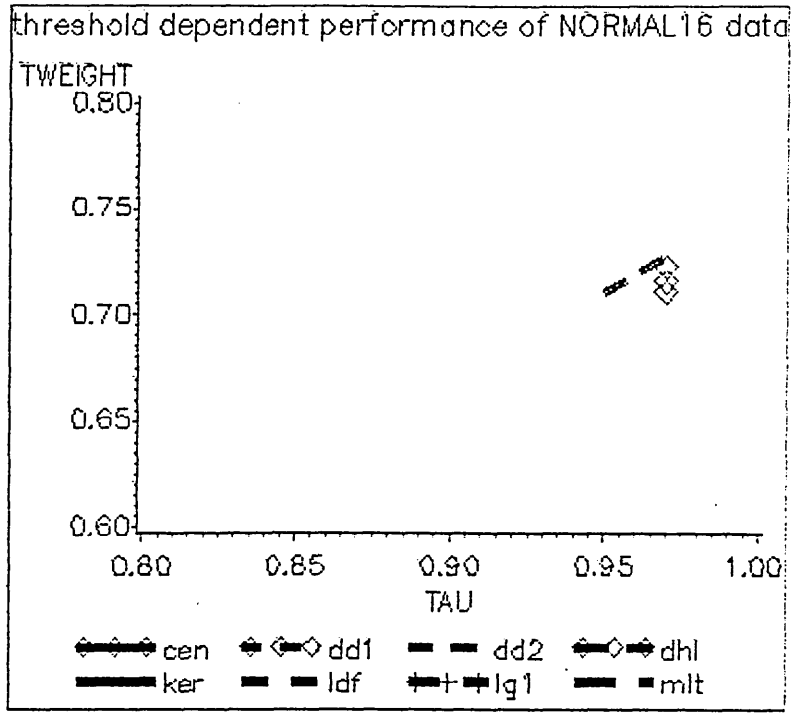


Figure G-11: Blow up of figure G-10

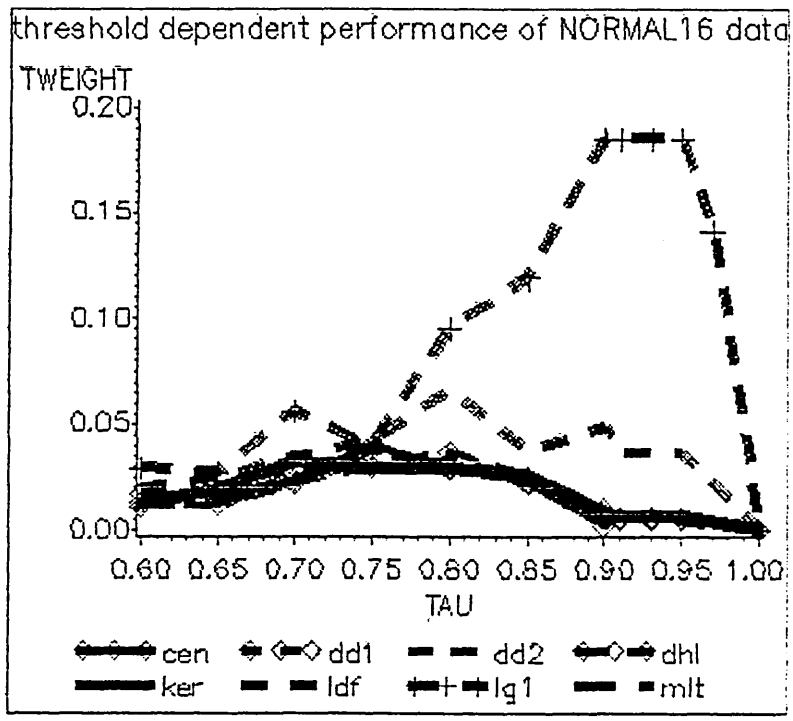


Figure G-12: Blow up of figure G-11

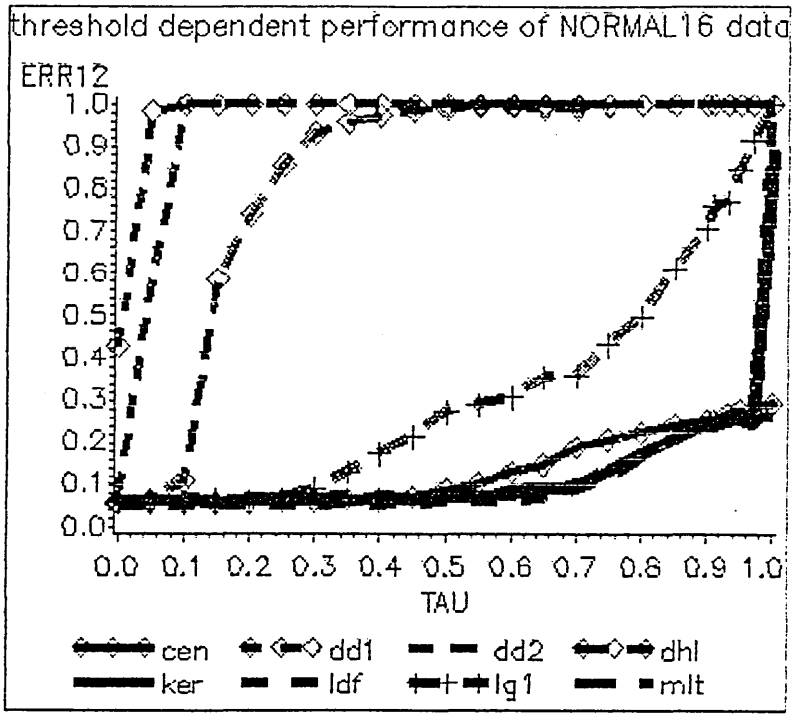


Figure G-13: Leave-1-out err(counting)

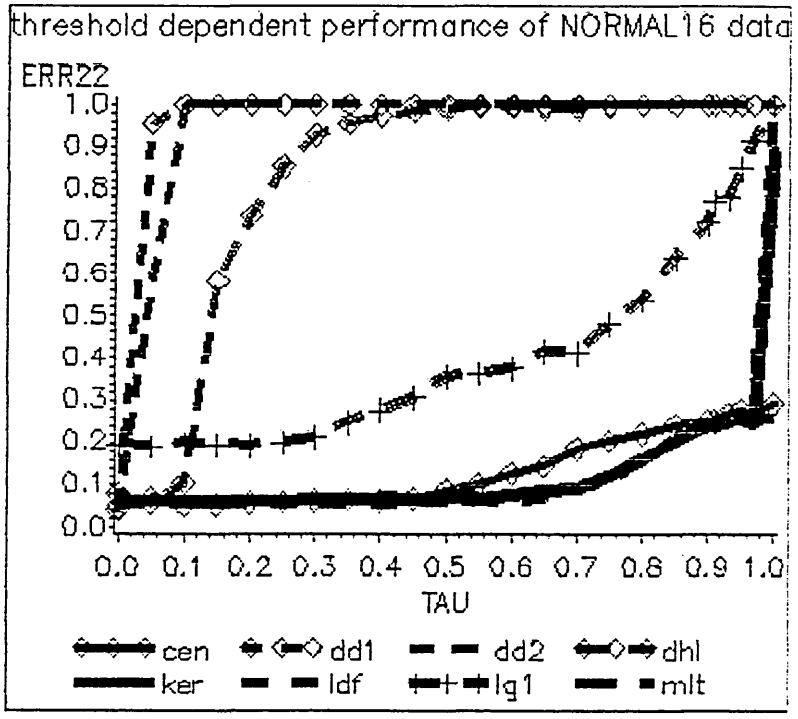


Figure G-14: Leave-1-out err(posterior)

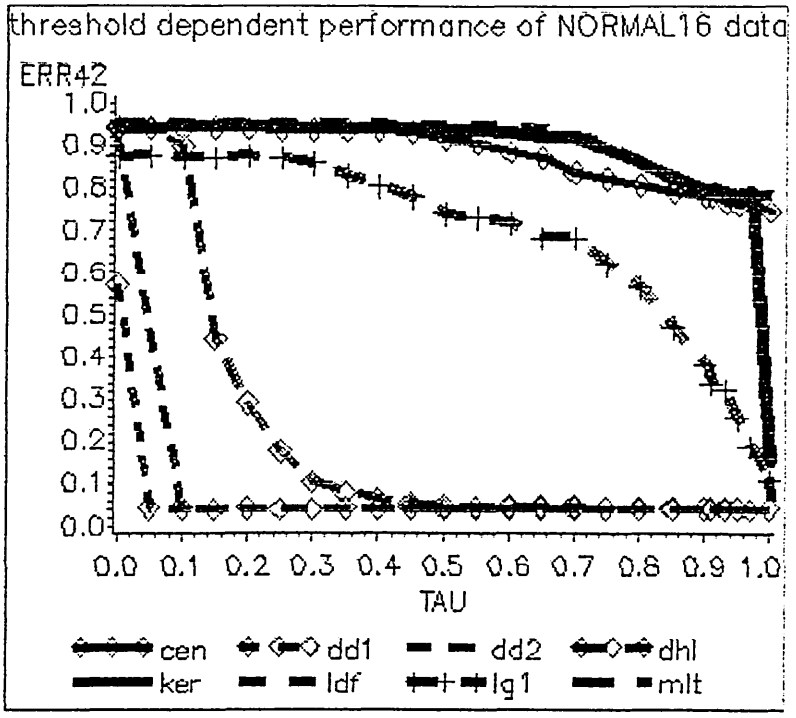


Figure G-15: Leave-1-out η

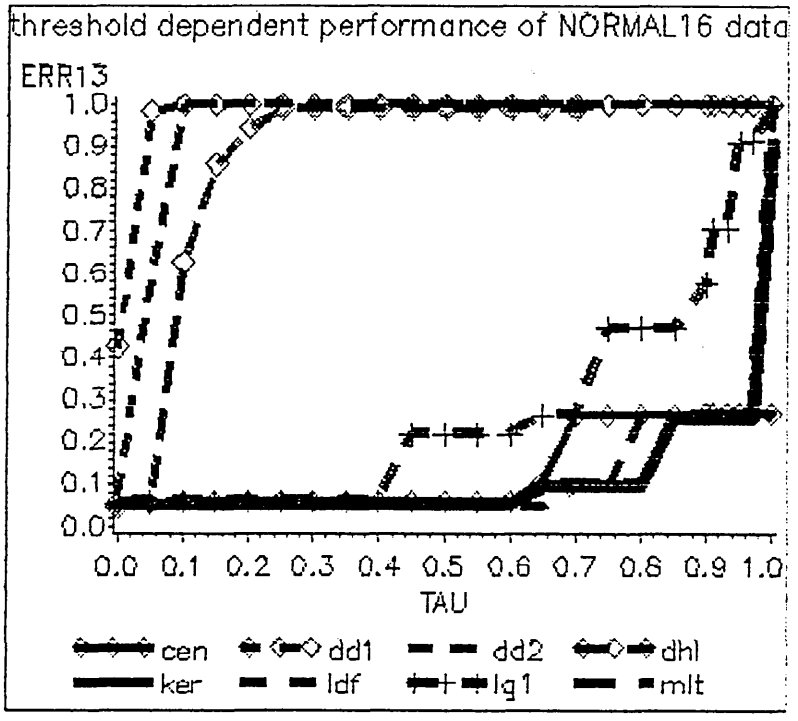


Figure G-16: Hold-out err(counting)

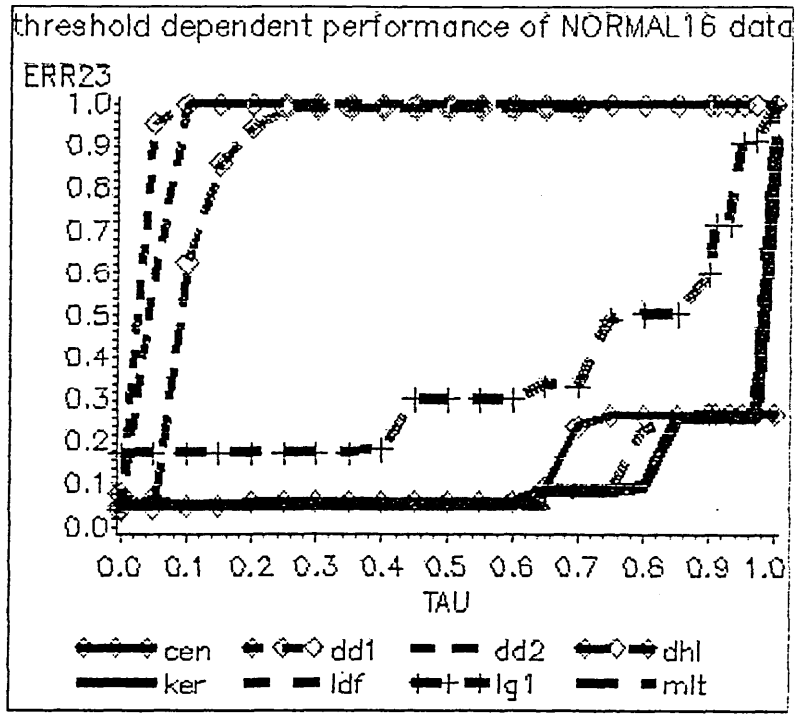


Figure G-17: Hold-out err(posterior)

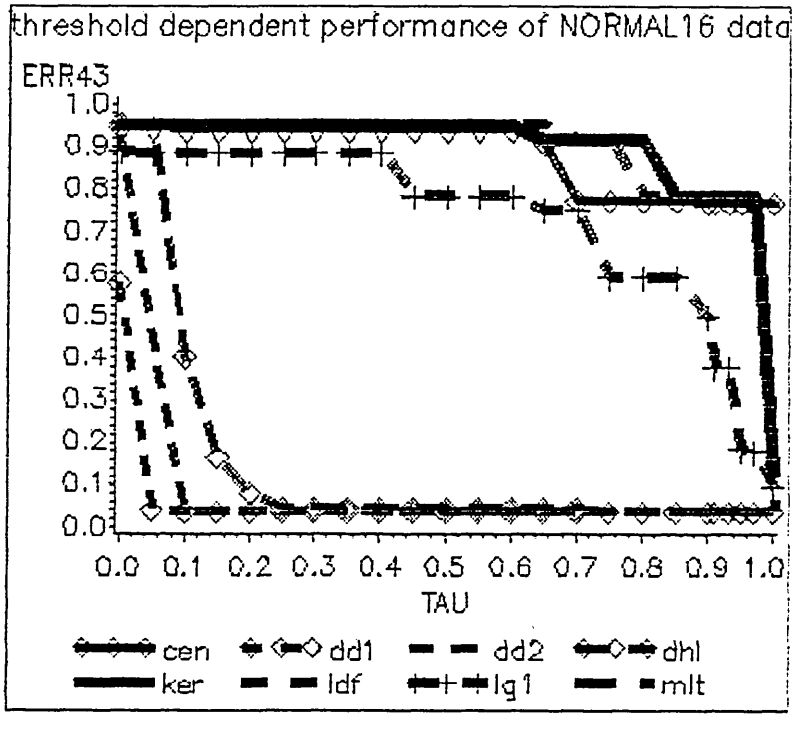


Figure G-18: Hold-out η

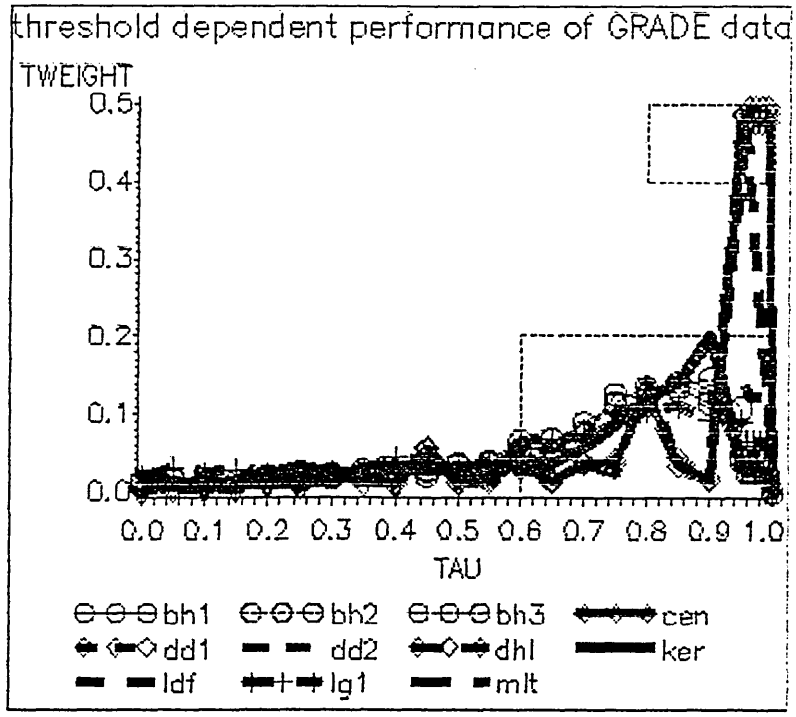


Figure G-19: Distribution of $f(\tau)$

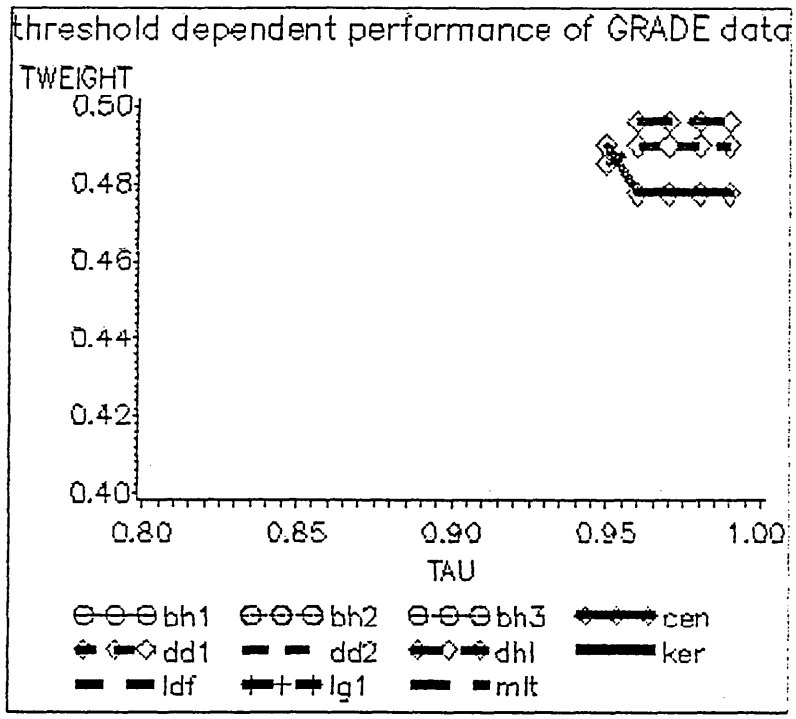


Figure G-20: Blow up of figure G-19

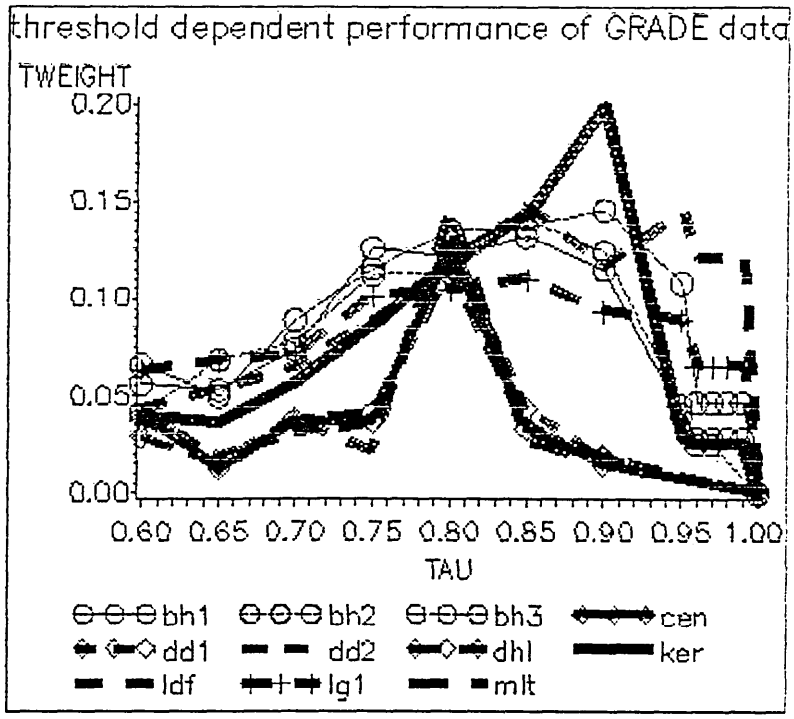


Figure G-21: Blow up of figure G-20

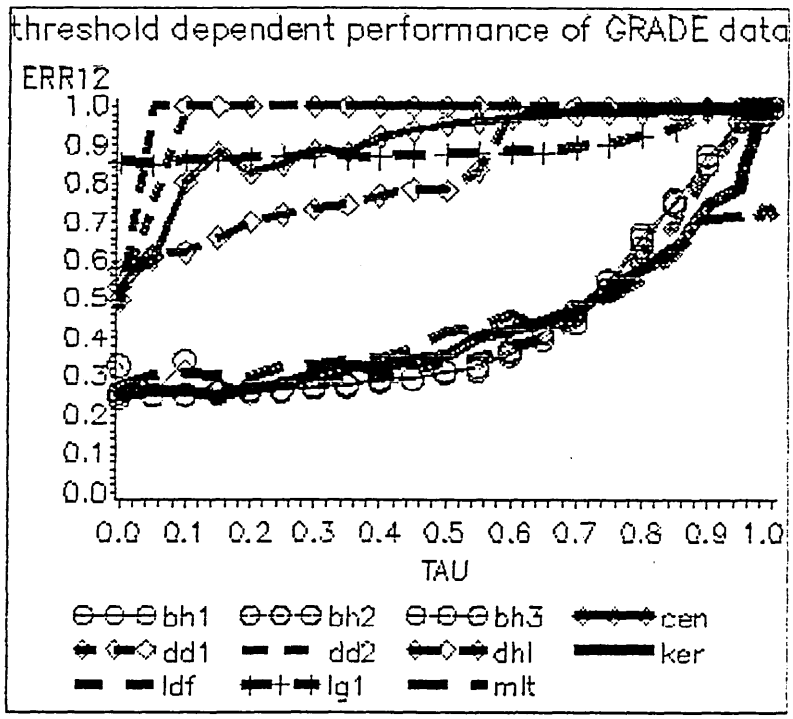


Figure G-22: Leave-1-out err(counting)

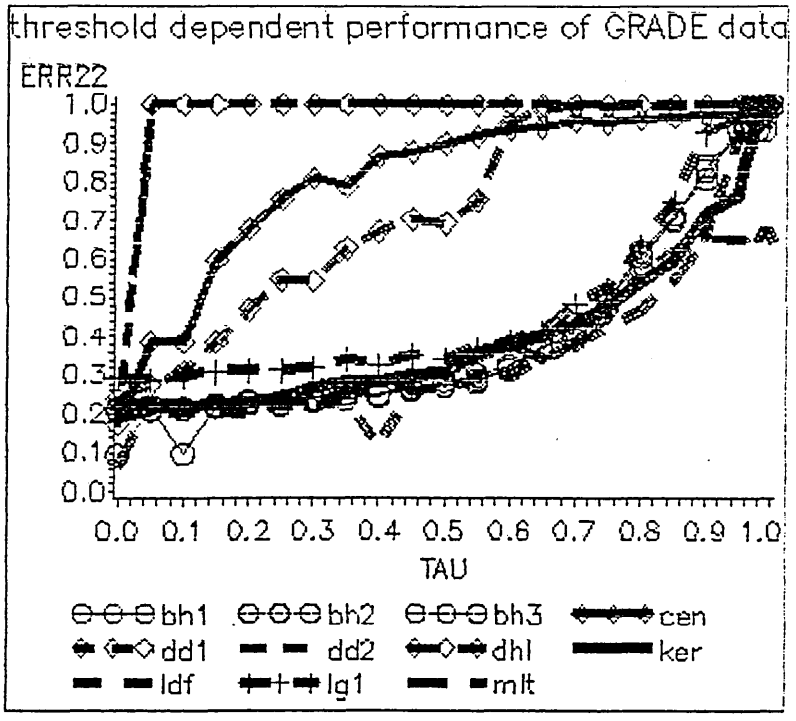


Figure G-23: Leave-1-out err(posterior)

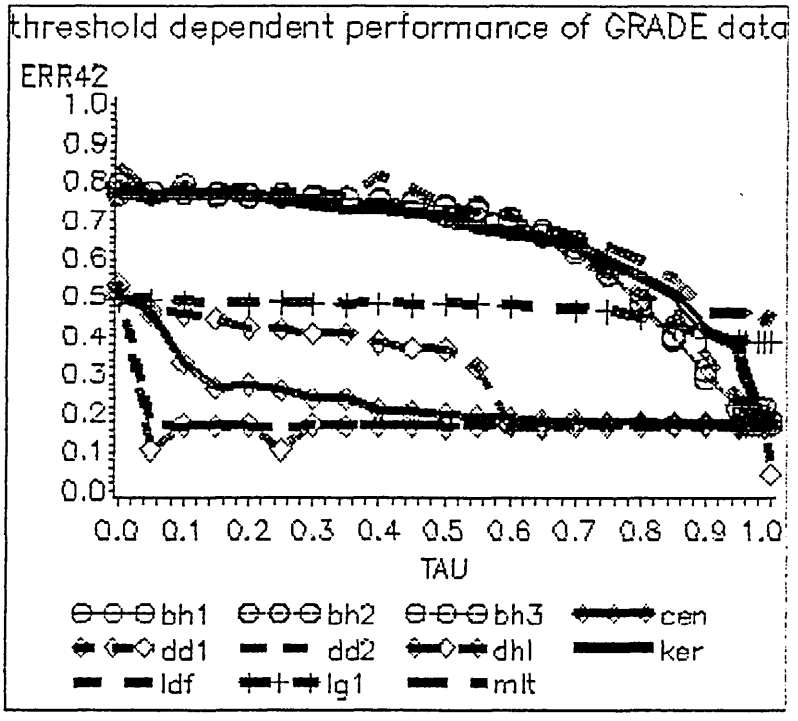


Figure G-24: Leave-1-out η

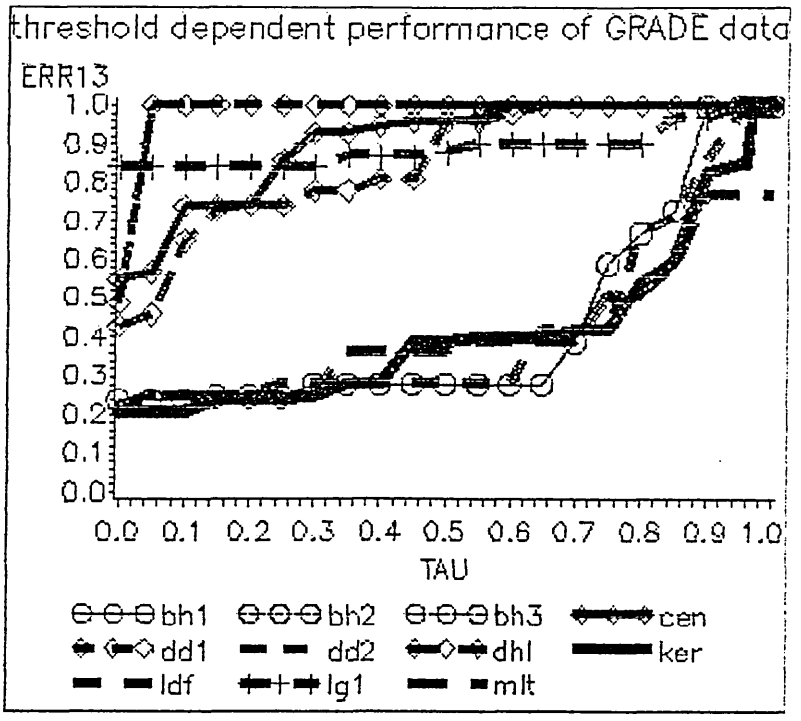


Figure G-25: Hold-out err(counting)

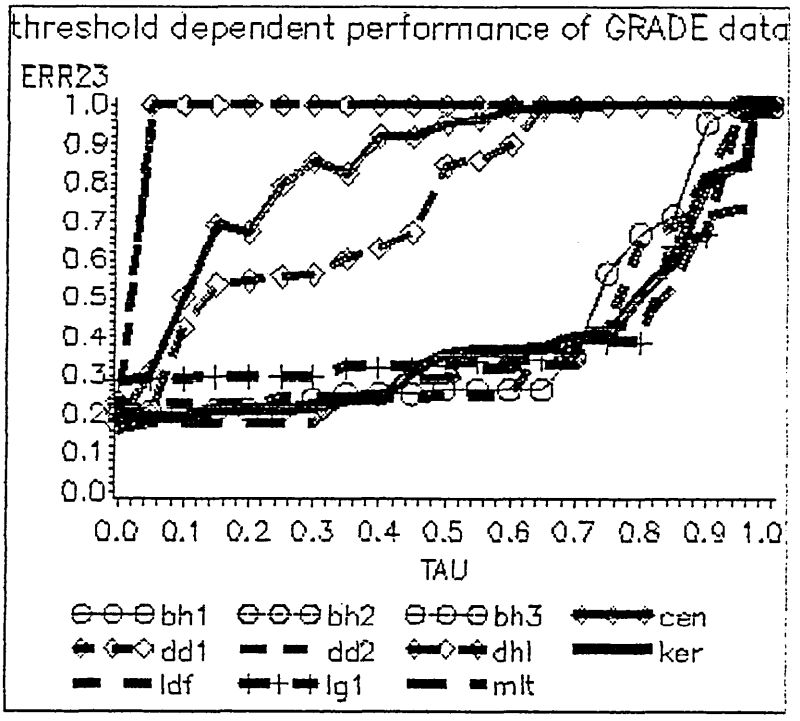


Figure G-26: Hold-out err(posterior)

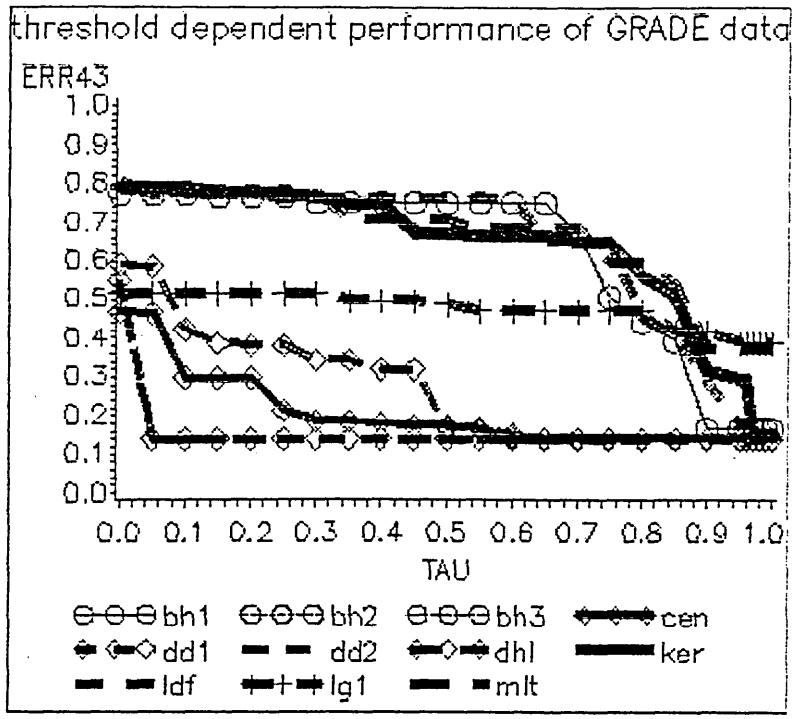


Figure G-27: Hold-out η

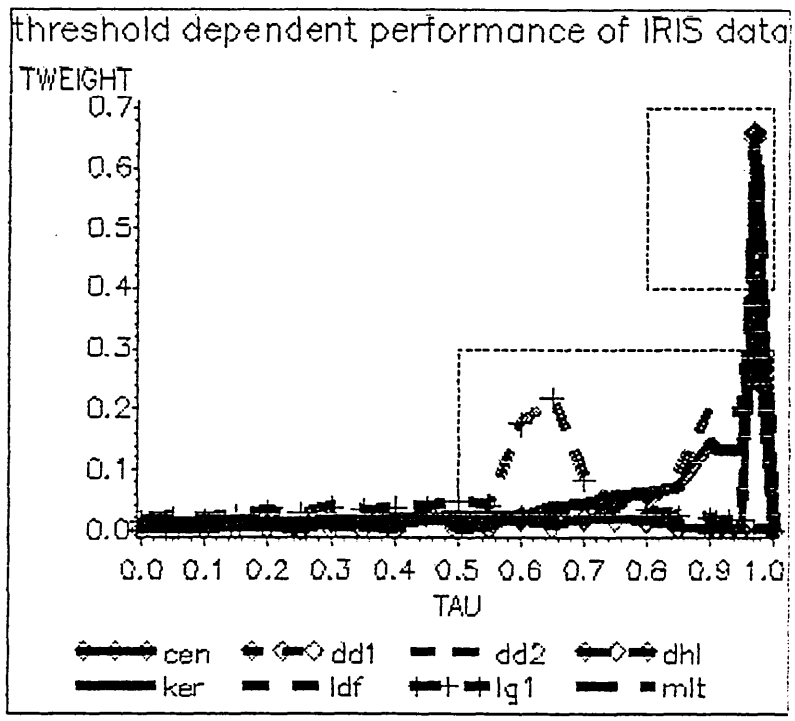


Figure G-28: Distribution of $f(\tau)$

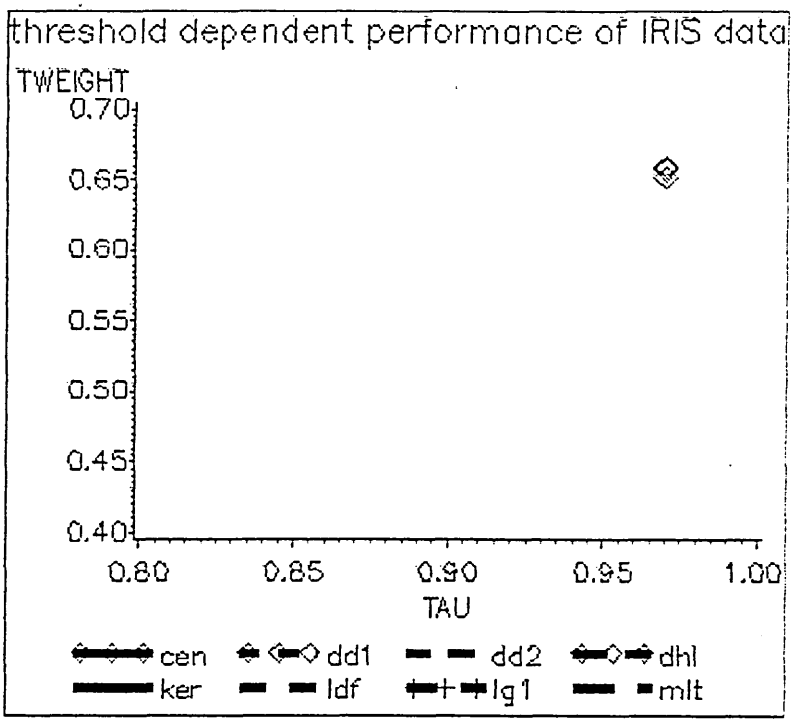


Figure G-29: Blow up of figure G-28

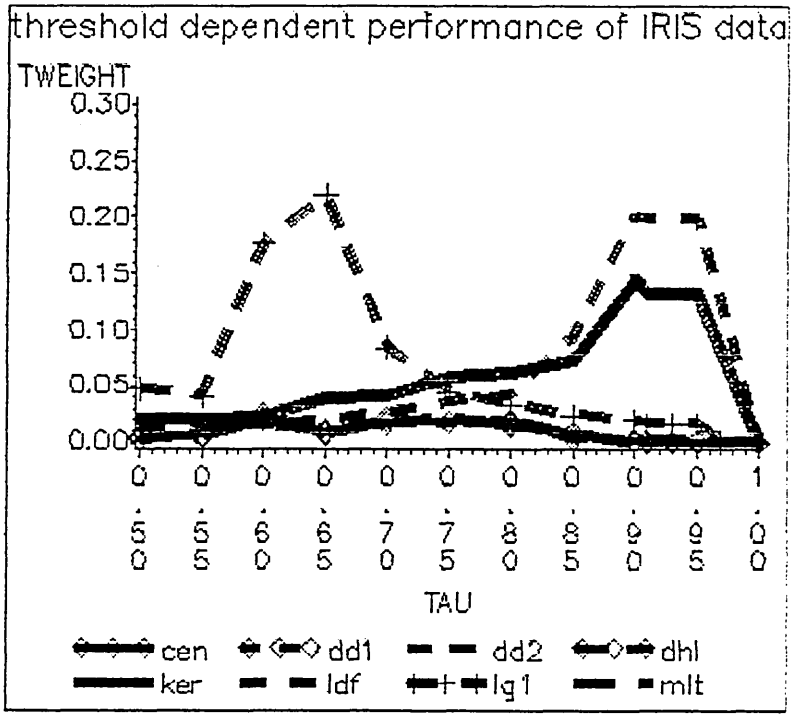


Figure G-30: Blow up of figure G-29

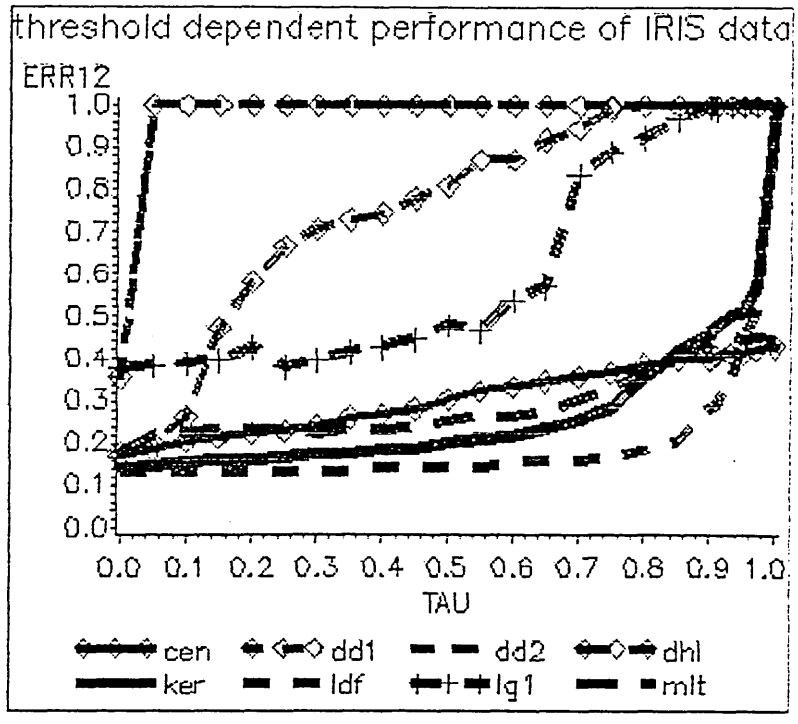


Figure G-31: Leave-1-out err(counting)

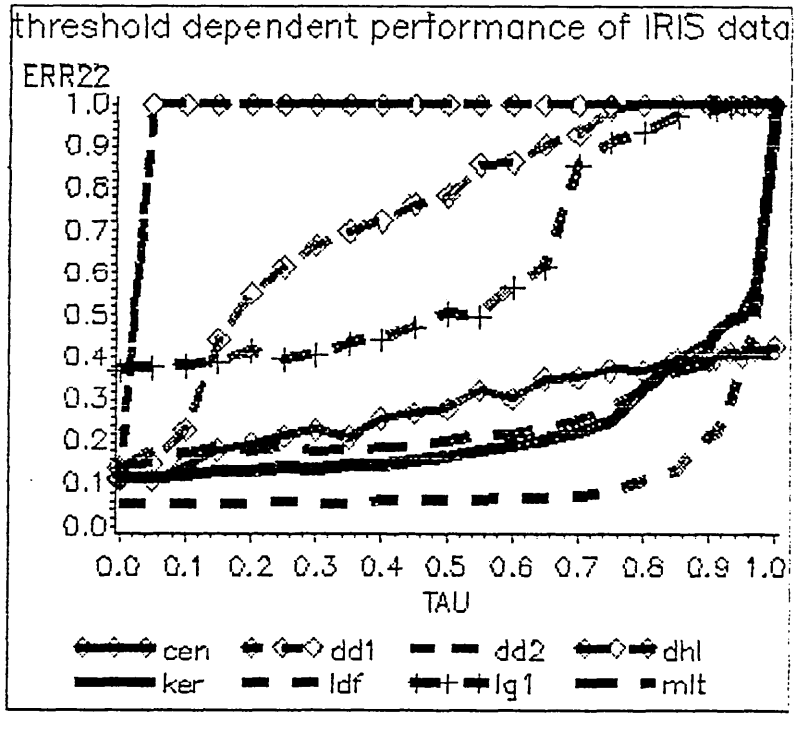


Figure G-32: Leave-1-out err(posterior)

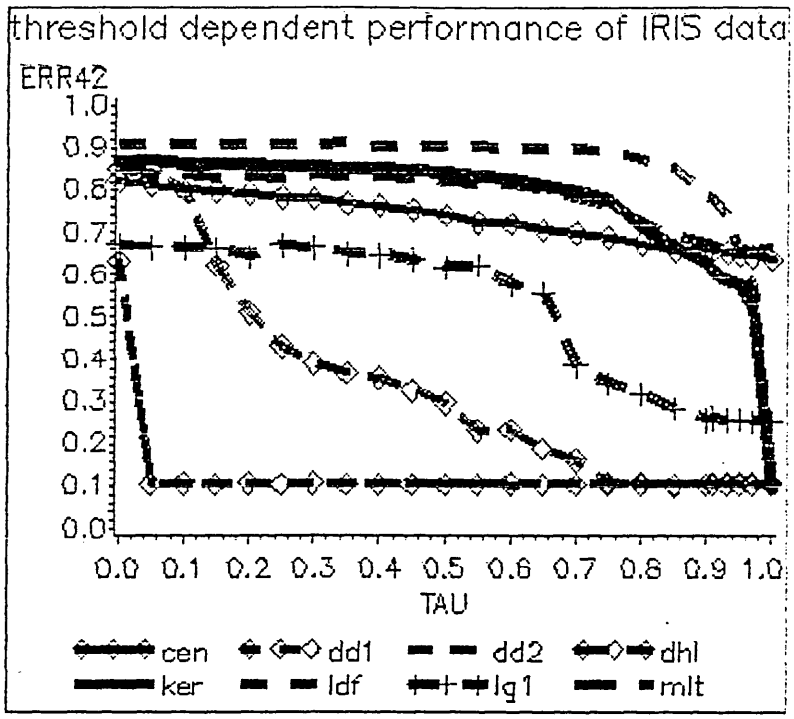


Figure G-33: Leave-1-out η

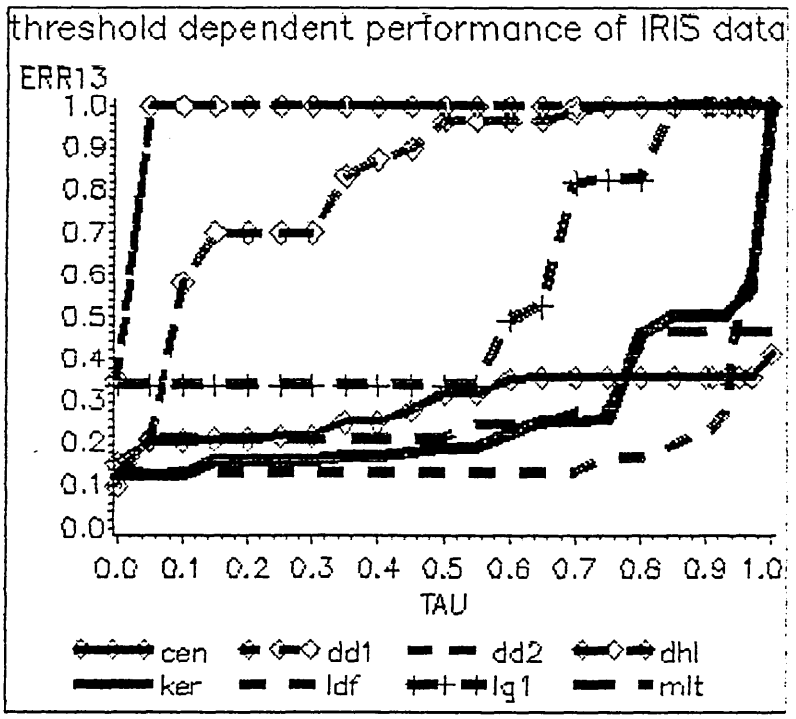


Figure G-34: Hold-out err(counting)

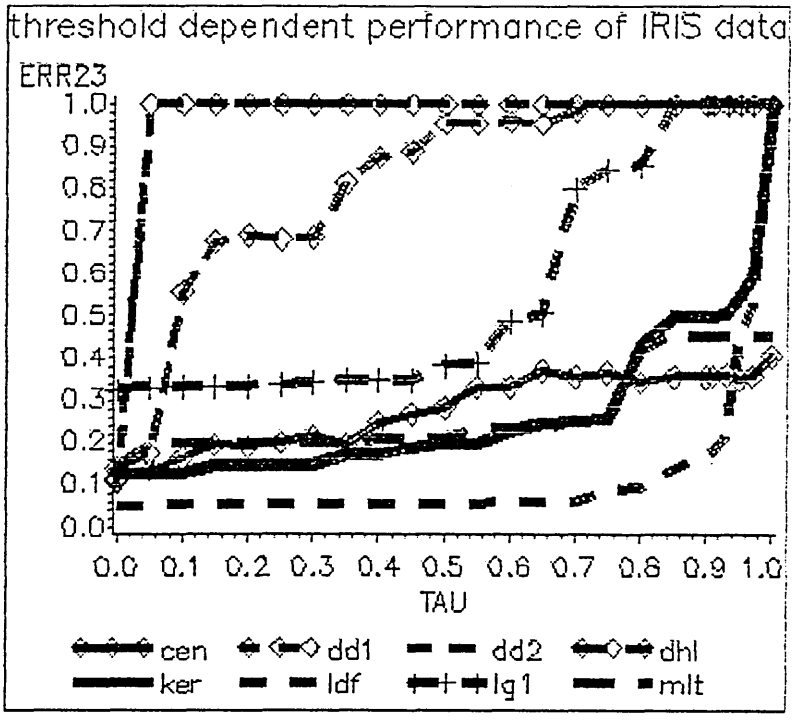


Figure G-35: Hold-out err(posterior)

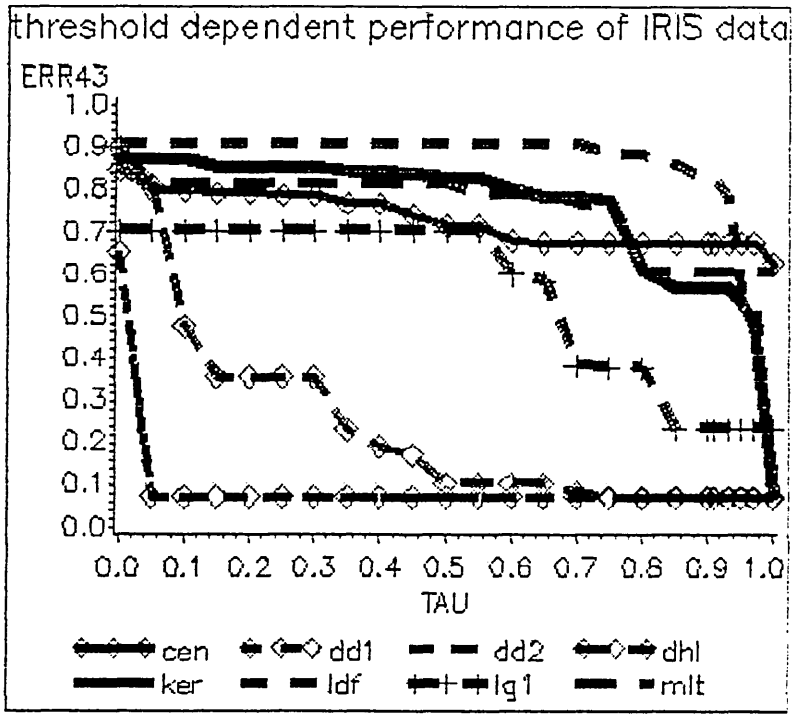


Figure G-36: Hold-out η